Competing Neural Networks as Models for Non Stationary Financial Time Series

- Changepoint Analysis -

Tadjuidje Kamgaing, Joseph

Vom Fachbereich Mathematik der Technische Universität Kaiserslautern
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Doctor rerum naturalium, Dr. rer. nat.)
genehmigte Dissertation

1. Gutachter: Prof. Dr. Jürgen Franke
2. Gutachter: Prof. Dr. Michael H. Neumann

VOLLZUG DER PROMOTION: 14. FEBRUAR 2005

D 386

To my family.

Acknowledgment

I am profoundly grateful to my supervisor, Prof. Jürgen Franke. He provides me with the topic, supports and encourages me along the way. On a personal level, I am deeply thankful for the confidence he place in me. Further, I thank Prof. Michael H. Neumann who accepts to be the second advisor for my thesis.
I would also like to thank Prof. Ralf Korn and through him the entire department of finance at the Fraunhofer ITWM (Institute for industrial mathematics) in Kaiserslautern where I was provided an office, a friendly and creative atmosphere as well as support to carry out my research.In particular, I thank all the people who shared the office with me during my thesis for the kindly, friendly and creative atmosphere.
I am also grateful to Prof. Marie Huskova(Charles University, Prague) for the introductory discussion on test in changepoint analysis we had during her visit last September at the university of Kaiserslautern.
I am deeply indebted to Dr. Jean Pierre Stockis and Dr. Gerald Kroisandt for their useful critics and the fruitful scientific discussion we use to have. Furthermore, I also deserve my gratitude to the entire Statistics research group of the university of Kaiserslautern for the friendly atmosphere and particularly, I deserve great respect to the secretary Mrs. Beate Siegler for this continuous achievement. Moreover, the funding of the Fraunhofer ITWM and Forschungsschwerpunkt Mathematik & Praxis of the mathematics department are highly appreciated.
Last but not least, I am thankful to my family and friends for their permanent support and to Elsy for her patience.
May God bless and continue to inspire all the people I mentioned above and those I silently and respectfully carry in my heart.

Abstract

The problem of structural changes (variations) play a central role in many scientific fields. One of the most current debates is about climatic changes. Further, politicians, environmentalists, scientists, etc. are involved in this debate and almost everyone is concerned with the consequences of climatic changes.
However, in this thesis we will not move into the latter direction, i.e. the study of climatic changes. Instead, we consider models for analyzing changes in the dynamics of observed time series assuming these changes are driven by a non-observable stochastic process. To this end, we consider a first order stationary Markov Chain as hidden process and define the Generalized Mixture of AR-ARCH model(GMAR-ARCH) which is an extension of the classical ARCH model to suit to model with dynamical changes.
For this model we provide sufficient conditions that ensure its geometric ergodic property. Further, we define a conditional likelihood given the hidden process and a pseudo conditional likelihood in turn. For the pseudo conditional likelihood we assume that at each time instant the autoregressive and volatility functions can be suitably approximated by given Feedfoward Networks. Under this setting the consistency of the parameter estimates is derived and versions of the well-known Expectation Maximization algorithm and Viterbi Algorithm are designed to solve the problem numerically. Moreover, considering the volatility functions to be constants, we establish the consistency of the autoregressive functions estimates given some parametric classes of functions in general and some classes of single layer Feedfoward Networks in particular.
Beside this hidden Markov Driven model, we define as alternative a Weighted Least Squares for estimating the time of change and the autoregressive functions. For the latter formulation, we consider a mixture of independent nonlinear autoregressive processes and assume once more that the autoregressive functions can be approximated by given single layer Feedfoward Networks. We derive the consistency and asymptotic normality of the parameter estimates. Further, we prove the convergence of Backpropagation for this setting under some regularity assumptions.
Last but not least, we consider a Mixture of Nonlinear autoregressive processes with only one abrupt unknown changepoint and design a statistical test that can validate such changes.

Contents

Some Abbreviations and Symbols

Abbreviations

$\lim$	limit
$\max$	maximum
$\min$	minimum
$\sup$	supremum
i.i.d.	independent identically distributed
M.C.	Markov Chain

Symbols

$\exp(x)$	$= e^x$
$\lvert x \rvert$	Absolute value of x
$\lVert \theta \rVert$	Norm of the vector θ
$\mathbb{Z}$	$= \{\cdots, -2, -1, 0, 1, 2, \cdots\}$
$\mathbb{N}$	$= \{0, 1, 2, \cdots\}$
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	d-dimensional Euclidian space
$\mathbb{P}(\ A\ )$	Probability of the set A
$\mathbb{P}(\ A \mid B\ )$	Conditional probability of the set A given the set B
$\mathbb{E}(\ X\ )$	Expectation of the random variable X
$\mathbb{E}(\ X \mid A\ )$	Conditional expectation of the random variable X given the information contained in A
$\mathcal{N}(0, 1)$	Standard Normal distribution
$\mathcal{N}(0, \Sigma)$	Multivariate Normal distribution with mean vector 0 and covariance matric Σ

1 Introduction

1.1 Motivations

In various fields one has to analyze data collected over long periods of observation. Time series models account for one of the most widely used tools in data analysis. The classical time series behavior is to assume stationary stochastic processes as model for these data under the main hypothesis that these data satisfy some stability conditions or invariance properties. This hypothesis is satisfied by many linear models that have now been intensively used for many decades. For example the first order autoregressive processes

$$Y_t = \alpha Y_{t-1} + \varepsilon_t,$$

for which $|\alpha| < 1$ and w.l.o.g the residuals ε_t are random variables with mean zero and unit variance, e.g. $\mathcal{N}(0,1)$. The following plot contains examples of such computer-generated processes for which we have considered $\alpha = 0.97$ and -0.97 respectively.

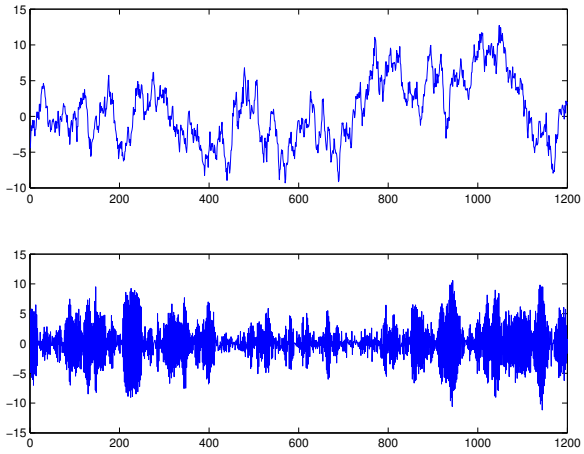

Figure 1.1: Stationary First Order Autoregressive Processes

However, these assumptions (e.g. of invariance in time) are frequently satisfied only over periods of limited length, in other words they are usually only locally satisfied as one can observe in some very specific cases. We can consider for example a simple mixture of two stationary first order autoregressive processes as illustrated by Figure 1.2. Under this setting the regular variation of the structure is clearly exhibited and human eyes can also be used to make the decision on such changes. Unfortunately, it is not always the case that this violation of the invariance property is clearly observable just by using human eyes as confirmed by Figure 1.3. In fact, in this picture it is less obvious than in the previous one where the changes may have

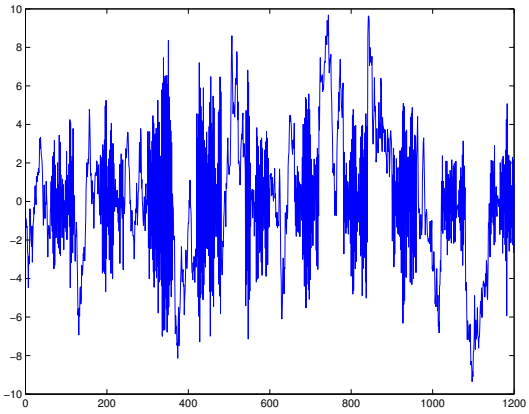

Figure 1.2: Mixture of Stationary AR(1)

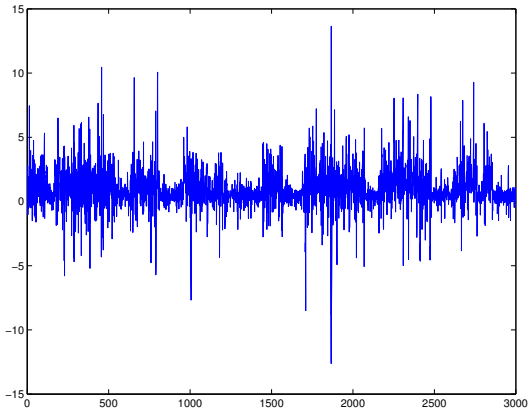

Figure 1.3: Mixture of NLAR-ARCH

occurred. At this point we just claim for Figure 1.3 that the invariance principle is indeed violated, which we will make clear later on.
The last two graphics illustrate problems that belong to a very broad class, that of detecting changes in the structure of a continually observed time series for which applications can be found in many fields, for example in finance, industrial quality control, medical sciences (monitoring of patients), speech recognition and meteorology.
In general, for this class of problem, one will face three main types of situations: the changes in the mean, the change in the variability and the change in the dependence structure of the process. In the current work we focus on the latter situation and propose a quite general time series model which repeatedly moves from one state to a different state. Moreover, we discuss two algorithms which, after a period of initialization, are able to detect these change-points.

1.2 Outline

The aim of the current work is to develop new models and algorithms which enable the modeling of time series under the assumption of change in the dependence structure of the observed processes.
For this aim, we extend the class of ARCH models (introduced in 1982 by Engle) to a more general class of models, namely the generalized mixture of nonlinear AR-ARCH models that is presented in Chapter 2. In this chapter, the models assumptions are presented, the first two conditional moments, the conditional distribution and the conditional likelihood are derived in turn under special conditions. Moreover, the geometric ergodicity, i.e. the asymptotic stability property of such models is established under more general considerations.
In Chapter 3 we define a nonlinear conditional least squares approach. Following an

idea of Franke et al we prove the asymptotic consistency of the autoregressive function estimates given some parametric classes of functions. This result is particularly valid for some classes of feedfoward network functions.
As alternative to the conditional nonlinear least squares defined in Chapter 3, in Chapter 4 (based on the hidden process) we define a conditional likelihood from which we derive a pseudo conditional $\log$-likelihood. Indeed, for the pseudo conditional $\log$-likelihood we assume that the autoregressive and volatility functions can be suitably approximated by feedfoward networks with a fixed number of hidden neurons. Under this setting the consistency of the parameter estimates is proven for the pseudo conditional $\log$-likelihood. Moreover, a version of the Expectation-Maximization (EM) Algorithm introduced by Baum et al is proposed to solve the problem numerically.
In Chapter 5, focusing on the changes driven by the autoregressive functions, we propose some weighted least squares techniques for estimating the changes in the dynamics of the observed process. Under the assumption that the autoregressive function can be suitably approximated by feedfoward network and under some regularity assumptions, the consistency and asymptotic normality of the parameter estimated are proven. The results of this chapter are extended in Chapter 6, where we assume a multivariate time series. Furthermore, in Chapter 7, following an idea by White [88], we prove the convergence and asymptotic normality of Backpropagation (a stochastic approximation algorithm that can be used to solve the problem numerically considering the weighted least squares).
Chapter 8 gives a short introduction to the problem of test in changepoint analysis. In fact, a nonlinear autoregressive case with only one abrupt unknown change is considered and a test for validating the change is designed under some regularity assumptions.
In chapter 9 some numerical applications of the the pseudo conditional $\log$-likelihood or hidden Markov techniques (developed in Chapter 4) and weighted least squares techniques (developed in Chapter 5) are presented. Indeed we present the results for the computer-generated data and real-life financial data as well.
The current work is summarized in Chapter 10 where some open questions of particular interest are exhibited. Finally, a brief introduction to Neural Networks is presented in Appendix A.

2 Generalized Nonlinear Mixture of AR-ARCH

2.1 Introduction

Since many decades time series models have been intensively used for analyzing the dynamic behavior of medical, social, economic, financial variables, etc. The most popular choices are linear models as autoregressive (AR), Moving Average (MA) and Mixed Autoregressive Moving Average (ARMA) processes. The linear time series models became very popular essentially because of their theoretical tractability and partly because they have been incorporated into many standard statistical software packages. Despite their popularity, linear models suffer from several drawbacks. These include their inability to capture dynamic patterns such as asymmetry and volatility clustering, just to name a few.
In the last decades, many nonlinear time series models with successful applications have been proposed. We can mention, for example, the Autoregressive conditional heterokedastic models introduced in 1982 [23] by Engle (the Nobel Laureate) to model financial volatility. Additionally, we can refer to the book by Tong [84] for a general introduction on the nonlinear time series models. However, the nonlinear time series models also have their limitations. The main drawback is that most of the nonlinear models are designed just to describe specific nonlinear pattern, i.e. they may suffer from a lack of flexibility. Therefore a nonlinear model will be successful only if it is applied to a very specific class of data. An exception in this class is the so-called non parametric Artificial Neural network, that with its "Universal approximation property" is able to capture any nonlinear pattern into the data. Nevertheless, Neural Network can suffer from identifiability problem and therefore may also be vulnerable. In spite of the universal approximation property neural network may locally be inefficient if we consider very complex dynamical structures that for example exhibit local instability.
Since the end of the $1980s$ the Hidden Markov model (or Switching Markov model) in the framework of Hamilton [41] (who did some pioneer applications of this model to the US gross domestic product-GDP- growth) is gaining popularity. These models consist of different sub-models that can account for the behavior of the observed data in different dynamics. By allowing switches between these different sub-models, the new model (called mixture model) is able to represent more complex dynamical systems. Therefore, this models provides more flexibility than classical linear and nonlinear models. In literature one can find quite a lot of publications on this topic see, e.g. Hamilton ([41], [42], [43]), Elliot et al [22], Macdonald et al [64], Wong and Li [94], Stockis et al [83]. In general, the related publications contain a lot about the practical applications of the models but very few on their statistical properties. In this section we will consider a class of Generalized Nonlinear Mixture of AR-ARCH for which we will give a mathematical description, derive some of their basic properties and finally present and prove some results on their geometric ergodicity.

2.2 Model Description

In this section we present a general description of our model and give some assumptions that will be used in this chapter. Let us first present some definitions.

Definition 2.2.1 *Let us consider the hidden stochastic process* $\{Q_t,\ t \in \mathbb{N}\}$ *that takes its value on* $I_K = \{1, \cdots, K\}$*, where* K *is a given positive integer. Let us now define for* $k \in I_K$ *the stochastic processes*

$$S_{t,k}(\omega) = \begin{cases} 1 & \textit{if } k = Q_t(\omega) \\ 0 & \textit{otherwise}, \end{cases} \tag{2.1}$$

and in turn define the process $S_t(\omega) = (S_{t,1}(\omega), \cdots, S_{t,K}(\omega))$ *on* $\mathcal{K} \subset \mathbb{R}^K$ *the set of all its possible realizations.*

More details on the stochastic nature of Q_t will be given as one will need them, for example we will start by assuming the hidden stochastic process to be a stationary Markov Chain.
Consider a time series $\{X_t, \quad t = 0, 1, 2, \cdots\}$. For this series, we will assume that the underlying process has some changes in its dynamics. These changes can be modeled via a Hidden Markov Chain. This type of situation can be modeled with the help of a Generalized Mixture of AR-ARCH (GMAR-ARCH), i.e. a model defined as it follows.

Definition 2.2.2 *Generalized Mixture of AR-ARCH (GMAR-ARCH)*
A stochastic process is called a GMAR-ARCH of order K *and* p *if*

$$X_t = \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_{t,k}) \textit{ with } S_{t,k} = \begin{cases} 1 & \textit{for } k = Q_t \\ 0 & \textit{otherwise}, \end{cases} \tag{2.2}$$

where the processes $\{\epsilon_{t,k}\}$ *are i.i.d. random variables, independent of* $\mathbb{X}_{t-1} = (X_{t-1}, \cdots, X_{t-p})$*; and mutually independent for each* $k \neq j = 1, \cdots, K$*, and w.l.o.g.* $\mathbb{E}\epsilon_{t,k} = 0, \quad \mathbb{E}\epsilon_{t,k}^2 = 1$*. The* $m_k(u)$ *and* $\sigma_k(u)$ *are unknown real valued functions.*

2.2.1 Some Classical Cases

From the model defined in 2.2.2 one can derive some classical and well-known special cases. In this light let us consider several situations.
If we were to consider the special case where $K = p = 1$ and were given

$$m(x) = ax \quad \text{and} \quad \sigma(x) = c > 0,$$

the process simply describes an AR (1). Similarly, one can represent a classical ARCH (1) model by choosing

$$m(x) \equiv 0 \text{ and } \sigma(x) = \sqrt{\omega + \alpha x^2}, \quad \omega > 0, \quad \alpha \geq 0.$$

However, if

$$m(x) = ax \quad \text{and} \quad \sigma^2(x) = \sigma x^2,$$

the process is reduced to the discrete time version of a geometric Brownian Motion as considered by Black and (the Nobel laureate) Scholes for their well-known option pricing model.
Still, we have to observe that our model differs from this classical model, let us consider for example $K = 2$. If we then assume $S_{t,1} = 1$ for $t = 1, \cdots, \tau_0$ and $S_{t,2} = 1$ for $t = \tau_0 + 1, \cdots, n$, the model defined in equation 2.2 experiences a single structural change as the parameters of the model abruptly change after τ_0. In the case where $K \geq 2$, we can, e.g. consider that $\{S_t\}$ as an i.i.d. sequence of random variables. Each dynamic variable is independent of the past and future dynamics and the process X_t may switch back and forth between different dynamics. That is the case under consideration in Quandt ([78], 1972) and the mixture of autoregressive, i.e. the model of equation 2.2 for which we assume that the $\{S_t\}$ are i.i.d. sequences of random variables, all volatility functions are constant and the m_k are linear, i.e.

$$m_k(u) = \sum_{i=1}^{p} \alpha_{k,i} u_i$$

as proposed by Wong and Li [94]. This is just a special case of such models.
These models are able to capture time series with several dynamics, but they suffer from different drawbacks. The model with one abrupt change is too restrictive in practice since it admits only one change. As time series are correlated by nature, it seems more convenient to expect that each dynamic depends on the past happenings. To overcome the limitations we can find in some classical models we can make various assumptions on the hidden process, e.g. we can assume it to be be a first order stationary Markov Chain. Let us now present some general assumptions for the model.

2.3 Model Assumptions

Let us consider the $\{\epsilon_{t,k}\}$ to be independent of $\mathcal{G}_{t-1} = \sigma\{X_r, r \leq t-1\}$, additionally, conditioned on the past information we also assume that S_t and the $\epsilon_{t,k}$ are uncorrelated. Moreover we assume

$$\mathbb{P}(Q_t = j \mid Q_{t-1} = i, Q_{t-1}, Q_{t-2} \cdots, \mathcal{G}_{t-1}) \;=\; \mathbb{P}(Q_t = j \mid Q_{t-1} = i). \quad (2.3)$$

The

$$m_k : \mathbb{R}^p \longrightarrow \mathbb{R} \quad \text{and} \quad \sigma_k : \mathbb{R}^p \longrightarrow (0, \infty)$$

are unknown functions which in general have to be estimated; K is the given number of states or dynamics in the process, p is the order of the underlying $NLAR - ARCH(p)$ processes

$$X_t = m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_{t,k},$$

with $\mathbb{X}_{t-1} = (X_{t-1}, \cdots, X_{t-p})$. However, p does not need to be the same for all the underlying $NLAR - ARCH$. Nevertheless, we can always assume it by taking $p = \max_k\{p_1, \cdots, p_k\}$ if we were to be given different orders for the underlying $NLAR - ARCH$ dynamics. Moreover we do not need the autoregressive and the volatility functions to depend on the same parameter, i.e. we can choose autoregressive functions of order p and volatility functions of order q as in Masry and Tjostheim [67].
As announced earlier, various assumption can be made on $\{Q_t,\ t = 0, 1, \cdots,\}$ defined on $\{1, \cdots, K\}$, e.g. one can assume it to be an irreducible and aperiodic Markov Chain (M.C.), i.e. a stationary M.C. with initial stationary distribution $\pi = (\pi_1, \cdots, \pi_k)$ and transition probability matrix $(a_{i,j})$. A short introduction on discrete Markov processes will be presented in Chapter 4 and for more general theory and approaches on Markov models we will refer to the books by Tong [84], Meyn et al [70] and Duflo [20] among others.

2.4 Basic Properties Derived from the Model

It is clearly observable that

$$S_t = (S_{t,1}, \cdots, S_{t,K}),$$

with its entries $S_{t,k}, k = 1, \cdots, K$ defined as in equation 2.2 , inherits the properties of Q_t. For example if we assume Q_t is M.C. with value on I_K, S_t will consequently be a M.C. on $\mathcal{K}$ for which one should have the following basic properties.

$$\begin{aligned} \mathbb{P}(S_{t,j} = 1 \mid S_{t-1,i} = 1) &= \mathbb{P}(Q_t = j \mid Q_{t-1} = i) \\ &= a_{i,j}, \end{aligned} \tag{2.4}$$

where the $a_{i,j}$ are the entries of transition probability matrix. We also have

$$\begin{aligned} \mathbb{E}S_{t,k} &= \mathbb{P}(Q_t = k) \\ &= \pi_k, \end{aligned} \tag{2.5}$$

with

$$\pi_1 + \cdots + \pi_K = 1 \tag{2.6}$$

since π is the initial stationary distribution of Q_t.
In the next two subsections we strengthen the Markov assumption on Q_t and simply assume the S_t to be a sequence of i.i.d. random variables. Under this simple assumption we give an impression of how one can compute the conditional expectation, conditional variance and conditional distribution for our model given the past information.

2.4.1 Conditional Moments

The purpose of this section is to compute first and second order conditional expectation of X_t given the past observations $\mathbb{X}_{t-1}$. Assuming S_t is a sequence of i.i.d. random variable it follows trivially that

$$\mathbb{E}(S_{t,k} \mid \mathbb{X}_{t-1} = x) \quad = \quad \pi_k.$$

Given this property, we can derive the conditional expectation and the conditional variance of X_t with respect to the past information $\mathbb{X}_{t-1}$, what we summarize in the following lemma.

Lemma 2.1 *Under the model assumptions and assuming S_t i.i.d. sequences of random variables it follows that the conditional expectation of X_t with respect to the past realizations $\mathbb{X}_{t-1}$ is defined as*

$$\mathbb{E}(X_t \mid \mathbb{X}_{t-1} = x) \quad = \quad \sum_{k=1}^{K} \pi_k m_k(x) \tag{2.7}$$

and its conditional variance is defined as

$$\begin{aligned} var(X_t \mid \mathbb{X}_{t-1} = x) \quad &= \quad \sum_{k=1}^{K} \pi_k \left(m_k^2(x) + \sigma_k^2(x) \right) \\ &\quad - \left(\sum_{k=1}^{K} \pi_k m_k(x) \right)^2 . \end{aligned} \tag{2.8}$$

Remark 2.2 *For the conditional variance, let us observe that*

$$\sum_{k=1}^{K} \pi_k m_k^2(x) - \left(\sum_{k=1}^{K} \pi_k m_k(x) \right)^2$$

is non negative since the square function is convex. It takes the value zero if we consider $m_j(x) = m_k(x)$ for all $j, k = 1, \cdots, K$. Therefore, we can derive the smallest value of the volatility, which we can write as follows

$$\sum_{k=1}^{K} \pi_k \sigma_k^2(x).$$

We can consider the latter as baseline for the volatility at each time instant.

Let us now present a proof of the above lemma.
Proof: By definition, the conditional expectation can be written as

$$\begin{aligned}\mathbb{E}(X_t \mid \mathbb{X}_{t-1} = x) &= \sum_{k=1}^{K} m_k(x)\mathbb{E}(S_{t,k} \mid \mathbb{X}_{t-1} = x) \\ &+ \sum_{k=1}^{K} \sigma_k(x)\mathbb{E}(S_{t,k}\epsilon_t \mid \mathbb{X}_{t-1} = x).\end{aligned}$$

For the first part of this equation we need to apply the result from equation 2.5 and for the second part we use the fact that S_t and ϵ_t are uncorrelated, conditioned on the past information.
For the proof of the conditional variance, it suffices to derive the conditional second moment, i.e.

$$\mathbb{E}(X_t^2 \mid \mathbb{X}_{t-1} = x) = \sum_{k=1}^{K} \pi_k(m_k^2(x) + \sigma_k^2(x)).$$

To prove it let us remark that

$$X_t^2 = \sum_{k=1}^{K} S_{t,k}(m_k^2(\mathbb{X}_{t-1}) + 2\epsilon_t m_k(\mathbb{X}_{t-1})\sigma_k(\mathbb{X}_{t-1}) + \epsilon_t^2\sigma_k^2(\mathbb{X}_{t-1}))$$

and the proof is similar to that for the conditional expectation. The conditional variance is then obviously derived. ∎

Once we have derived the conditional moments, let us now derive the conditional distribution under mild assumptions.

2.4.2 Conditional Distribution

Since in the case of the mixture of time series the conditional distribution can be multi-modal, one has the feeling that the conditional mean may not be the best predictor of the future values of the series. However, the merits of our model partly find their justification in their ability to provide nice formula for the conditional distribution with respect to the past information. For sake of simplicity, if we then assume that the $\epsilon_{t,i} = \epsilon_{t,j} = \epsilon_t \; \forall \; i, j = 1, \cdots, K$, are i.i.d. normally distributed, the conditional distribution of X_t given $\mathbb{X}_{t-1} = x$ is given in the following lemma.

Lemma 2.3 *We consider the model assumptions, S_t i.i.d. random variables and the*

$$\epsilon_{t,i} = \epsilon_{t,j} = \epsilon_t \; \forall \; i, \; j = 1, \cdots, K,$$

i.i.d. standard normally distributed. Then, it follows that

$$F(X_t \mid \mathbb{X}_{t-1} = x) = \sum_{k=1}^{K} \pi_k \Phi\left(\frac{X_t - m_k(x)}{\sigma_k(x)}\right), \tag{2.9}$$

where Φ is the cumulative distribution of the standard normal distribution.

Proof: To prove it, let us first consider its conditional density, i.e.,

$$\begin{aligned} f(X_t \mid \mathbb{X}_{t-1} = x) &= \sum_{k=1}^{K} f(X_t, Q_t = k \mid \mathbb{X}_{t-1} = x) \\ &= f(X_t \mid Q_t = k, \mathbb{X}_{t-1} = x) f(Q_t = k \mid \mathbb{X}_{t-1} = x) \end{aligned}$$

where $f(Q_t = k \mid \mathbb{X}_{t-1} = x)$ denotes the conditional probability weights of Q_t given $\mathbb{X}_{t-1} = x$.
By an application of the Tower property and the definition of the conditional density, we obtain

$$\begin{aligned} f(Q_t = k \mid \mathbb{X}_{t-1} = x) &= f(S_{t,k} = 1 \mid \mathbb{X}_{t-1} = x) \\ &= \mathbb{E}(\mathbb{E}(S_{t,k} \mid \mathbb{X}_{t-1} = x, Q_{t-1}) \mid \mathbb{X}_{t-1} = x) \\ &= \pi_k \end{aligned}$$

by equation 2.5. The integration of the conditional density with respect to X_t concludes the proof of the conditional distribution. ∎

Having this representation one can define the conditional $\log$-likelihood as

$$l = \sum_t \log \left\{ \sum_{k=1}^{K} \pi_k \Phi \left(\frac{X_t - m_k(x)}{\sigma_k(x)} \right) \right\}. \tag{2.10}$$

However, we have to remark that, even under i.i.d. assumption on the S_t, a direct computation of this quantity (l) may be computationally too demanding. This is partly because we have to compute the logarithm of a sum. To overcome this type of problem, we will propose an alternative solution for the computation of the conditional $\log$-likelihood in a later chapter by making use of the hidden structure.

2.5 Geometric Ergodicity

In time series analysis, we are generally interested in the stability of the model, i.e., in ergodicity. This is partly because of the theoretical importance of stationarity. However, it seems to appear that geometric ergodic models are very important since the rate of approaching stationarity ought to be fast enough for the stationary assumption to be relevant. Stability will therefore be the main purpose of this section in which we will find some sufficient conditions for determining whether the switching nonlinear autoregressive processes are geometric ergodic, what we will make clear in the coming sections.
Several authors have devoted works on stability conditions for nonlinear autoregressive processes, this is the case in Doukhan and Ghindes [19], Tong [84], Guégan and Diebolt ([38], 1994), Maercker ([66], 1995), Masry and Tjostheim [67] or more recently Z. Lu [63]just to name a few. In the switching setting we have some

work by Stockis et al ([83], on mixture of first order nonlinear AR-ARCH) and Yao et al [98] under some particular assumptions, among others.
Under weaker assumptions, we propose the geometric ergodic property of our process as defined in equation 2.2. For this purpose, we need to rely on $(S_t, X_t)'$ and therefore it will suffice to prove that $\zeta_t = (S_t, X_t, \cdots, X_{t-p+1})'$ is a geometric ergodic Markov Chain.

2.5.1 Assumptions, Markov and Feller Properties of the Chain

Before we move toward the geometric ergodic proof, let us first set down some assumptions and derive some properties that will help us to achieve our goal.

A. 2.4 *(Regularity Assumptions)*

1. *The process $\{Q_t\}$ is a first order stationary Markov Chain (S.M.C.) which is irreducible and aperiodic with initial stationary distribution $\pi = (\pi_1, \cdots, \pi_K)$ and transition probability matrix A.*
2. *The i.i.d. random variables ϵ_t have positive probability density function ϕ on $\mathbb{R}$ that is continuous, moreover they have zero expectation and finite variance; w.l.o.g. we will assume the latter equal to 1.*
3. *The functions m_k and σ_k are continuous on $\mathbb{R}^p$, and σ_k are bounded away from zero, i.e. $\inf\{\sigma_k(u) : u \in \mathbb{R}^p\} > 0 \quad \forall\, k \in \{1, \cdots, K\}$.*
4. *There exist $\alpha_k, d_k \in \mathbb{R}^p$ with $d_{k,i} \geq 0, i = 1, \cdots, p$, such that, as $\|u\| \to \infty$,*

$$m_k(u) = \sum_{i=1}^{p} \alpha_{k,i} u_i + o(\|u\|) \text{ and } \sigma_k^2(u) = \sum_{i=1}^{p} d_{k,i} u_i^2 + o(\|u\|^2).$$

As alternative to **A.**2.4-1, we can simply consider an irreducible and aperiodic chain and the stationary condition will follow directly, hence, we will avoid any redundancy.
Assumptions **A.**2.4-2 to 4 are analogous conditions due to Masry and Tjostheim [67] where they considered stability conditions for NAR-ARCH (p). The difference in our context is due to the GMAR-ARCH (p) with continuous autoregressive and volatility functions. Moreover, assumption A.2.4 4 does not imply that the model has to be parametric or linear in a certain sense. It just means that if the previous realizations of the observed process are large enough, one can prevent any explosion of the process to infinity by approximating the process with a parametric model as defined in the assumption.
Having stated the above assumptions (**A.**2.4), let us first state and prove that ζ_t is a Markov Chain.

Lemma 2.5 *Assuming the model assumptions and Q_t first order stationary M.C., it follows that ζ_t is a Markov Chain.*

Proof: To prove the above assertion, let us rewrite $\zeta_t = \binom{S_t}{U_t}$ with $U_t = (X_t, \cdots, X_{t-p+1})'$ and prove the claim.
Since S_t is a M.C., by Definition 6.1.3 in Duflo [20], there exists a measurable function F and a sequence η_t of identically distributed random variables, independent of ϵ_t, S_{t-1} such that

$$S_t = F(S_{t-1}, \eta_t).$$

Writing

$$U_t = \begin{pmatrix} X_t \\ . \\ . \\ . \\ X_{t+1-p} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_t) \\ X_{t-1} \\ . \\ . \\ X_{t+1-p} \end{pmatrix}$$

it follows that ζ_t is a Markov Chain with state space $\Omega \subset \mathcal{K} \times \mathbb{R}^p$. ∎

Given the previous lemma, we can now make sure that ζ_t satisfies some well-known and useful Markov properties.

Lemma 2.6 *Under the model assumptions and* **A.** *2.4 the chain $\{\zeta_t\}$ is irreducible and aperiodic. Moreover the chain has the Feller property, i.e. consider P as the transition probability kernel, then the mapping*

$$Ph(\zeta) = \int P(\zeta, dy)h(y), \quad \zeta \in \mathcal{K} \times \mathbb{R}^p$$

is bounded and continuous whenever h is a bounded, continuous function.

Proof: Consider $A \in \mathcal{B}, A \subseteq \Omega = \mathcal{K} \times \mathbb{R}^p$ with $\lambda(A) > 0$, ($\lambda = \nu\mu_p$ is the product of the counting measure ν on $I_K = \{1, \cdots, K\}$ and the Lebesgue measure μ_p on $\mathbb{R}^p$) and $\zeta_1 = \binom{S}{U_1}$. Using the technique by Tong [84] for the generalized nonlinear autoregressive model of order p that we extend to our model.
First, we remark that is it is sufficient to consider $A = \{s^*\} \times B$ for $s^* \in \mathcal{K}$ and $B \subseteq \mathbb{R}^p$ a Borel set, as, due to the finiteness of $\mathcal{K}$, any A can be written as finite disjoint union of such sets. Furthermore, choose k^* such that $S_j^* = 1$ for $j = k^*$, 0 else. We start with the case case $p = 1$, i.e. $X_t = U_t$. Then, the one-step transition probability give that we start in $\zeta_1 = (S_1, X_1)' = (s, x)'$ is with $S_j = 1$ for $j = l$, 0

else:

$$\begin{aligned}
&\mathbb{P}\left(\begin{pmatrix} S_2 \\ X_2 \end{pmatrix} \in A \mid S_1 = s, X_1 = x\right) \\
= \;& \mathbb{P}\left(Q_2 = k, X_2 \in B \mid Q_1 = l, X_1 = x\right) \\
= \;& \mathbb{P}\left(X_2 \in B \mid Q_2 = k, Q_1 = l, X_1 = x\right) \mathbb{P}(Q_2 = k | Q_1 = l, X_1 = x) \\
= \;& a_{l,k} \mathbb{P}(m_k(x) + \sigma_k(x)\epsilon_2 \in B) \\
= \;& a_{l,k} \int_B \frac{1}{\sigma_k(x)} \phi\left(\frac{u - m_k(x)}{\sigma_k(u)}\right) du \\
= \;& a_{l,k} b_k(x) \qquad (2.11)
\end{aligned}$$

For more than one step, we have analogous

$$\begin{aligned}
&\mathbb{P}\left(\begin{pmatrix} S_3 \\ X_3 \end{pmatrix} \in A \mid S_1 = s, X_1 = x\right) \\
= \;& \sum_{j=1}^{K} a_{l,j} a_{j,k} \int_B \int_{\mathbb{R}} \frac{1}{\sigma_k(y)} \phi\left(\frac{u - m_k(y)}{\sigma_k(y)}\right) \frac{1}{\sigma_j(x)} \phi\left(\frac{u - m_j(x)}{\sigma_j(u)}\right) dy du \\
= \;& \sum_{j=1}^{K} a_{l,j} a_{j,k} b_{j,k}(x) \qquad (2.12)
\end{aligned}$$

and doing so iteratively one obtains

$$\mathbb{P}\left(\begin{pmatrix} S_{t+1} \\ X_{t+1} \end{pmatrix} \in A \mid S_1 = s, X_1 = x\right) \;=\; \sum_{j_1, \cdots, j_t = 1}^{K} a_{l,j_1} \cdots a_{j_t,k} b_{j_1, \cdots, j_t, k}(x)$$

As ϕ is strictly positive and the $\sigma_k(x)$ are bounded away from 0 by **A.** 2.4, there is some $\delta_{x,t} > 0$ such that $b_{j_1, \cdots, j_t, k}(x) \geq \delta_{x,t} > 0$ for all $j_1, \cdots, j_t, k$, i.e.

$$\mathbb{P}\left(\begin{pmatrix} S_{t+1} \\ X_{t+1} \end{pmatrix} \in A \mid S_1 = s, X_1 = x\right) \;\geq\; \delta_{x,t} (A^t)_{l,k} > 0 \qquad (2.13)$$

for some t as the Markov chain Q_t is irreducible by **A.** 2.4. Therefore, A can be reached from any initial state $\binom{s}{x}$, and $\{\zeta_t\}$ is therefore irreducible and, analogously due to the aperiodicity of Q_t aperiodic.
For the case $p > 1$, the argument is essentially the same. One only has to take care of the fact that $U_{t,j} = U_{t-1,j-1},\ 2 \leq j \leq p$ such that usually an arbitrary set A with positive measure can be reached only after at least p steps.
Hence the chain $\{\zeta_t\}$ is irreducible and aperiodic. That $\{\zeta_t\}$ has a Feller property follows from the definition of the weak Feller property, the fact that the m_k, σ_k are continuous and the σ_k are bounded away from zero. ∎

In the coming sections we recall what it means for a stochastic process to satisfy the geometric ergodic property, present characterization and derive the property for the mixture of first order and higher GMAR-ARCH as well. Last but not the least, at the end of this chapter we derive some consequences of the asymptotic stability.

2.5.2 Asymptotic Stability and Small Sets

Let us first present two preliminary definitions. For this purpose, let us consider X_n Markov Chain on $(X, \mathcal{B}(X))$ with transition kernel P.

Definition 2.5.1 *Small set and petite set*

1. *A set $C \in \mathcal{B}(X)$ is called a small set if there exists an $m > 0$, and a non-trivial measure ν_m on $\mathcal{B}(X)$, such that for all $x \in C, B \in \mathcal{B}(X)$,*

$$P^m(x, B) \geq \nu_m(B)$$

2. *A set $C \in \mathcal{B}(X)$ is ν_a-petite if there exists a probability measure $a = \{a(n)\}$ on $\mathbb{N}$ such that*

$$\sum_{n=0}^{\infty} a(n) P^n(x, B) \geq \nu_a(B), \forall x \in C, B \in \mathcal{B}(X),$$

where ν_a is a non trivial measure on $\mathcal{B}(X)$.

From the above definitions, one can observe that a small set is obviously a petite set.

Lemma 2.7 *Let us assume $\{\zeta_t\}$ satisfies the conditions of Lemma 2.6. Suppose there exist a small set C, a non negative measurable function g, and constants $0 < r < 1, \quad \gamma > 0 \quad$ and $\quad B > 0$ such that*

$$\begin{aligned} &\mathbb{E}\left(g(\zeta_t) \mid \zeta_{t-1} = \binom{S}{x}\right) < rg\binom{S}{x} - \gamma, && \binom{S}{x} \notin C \\ &\mathbb{E}\left(g(\zeta_t) \mid \zeta_{t-1} = \binom{S}{x}\right) < B, && \binom{S}{x} \in C. \end{aligned}$$

Then the chain $\{\zeta_t\}$ is geometrically ergodic, i.e. ζ_t is ergodic with stationary probability distribution measure λ and there exists a positive constant $\rho < 1$ such that

$$\|P^n(. \mid \zeta) - \lambda\|_{TV} = O(\rho^n),$$

where

$$P^n(B \mid \zeta) = P(\zeta_n \in B \mid \zeta_o = \zeta), B \in \mathcal{B}$$

is the conditional distribution of ζ_n given $\zeta_o = \zeta$ and $\| \ \|_{TV}$ is the total variation norm.

Proof: Note that the assumptions of the lemma 2.6 ensure that the chain $\{\zeta_t\}$ is irreducible and aperiodic and the rest of the proof follows by using the fact that under this observation, the lemma is just a version of Theorem $A1.5$ of the book by Tong [84]. One can also refer, e.g. to Tweedie (1975) for the proof of this lemma. ∎

To make use of this lemma we essentially need to provide the function g and prove the existence of a small set. Let us first prove the existence of this small set.

Lemma 2.8 *Let **A**.2.4 hold, then every compact set is a small set.*

Proof: During the proof of this lemma, we will refer to Lemma 2.6 from this section and apply several properties from the book by Meyn and Tweedie [70]. Let us now consider ζ_t as irreducible chain for λ. By Proposition $4.2.2$ (i) there exists a maximal probability measure ψ for which ζ_t is ψ-irreducible. Let $A \subseteq \Omega = \mathcal{K} \times \mathbb{R}^p$ belongs to the σ-algebra defined on $\mathbb{R}^p \times \mathcal{K}$ with $\lambda(A) > 0$, then for all $x \in \Omega$ we have

$$L(x, A) = P_x(\zeta \quad \text{ever enters} \quad A) \geq P(x, A) > 0$$

as in the proof of Lemma 2.6. Hence, by Proposition $4.2.2$ (iii) $\psi(A) > 0$, i.e. the interior of the support of ψ is nonempty. Therefore, by Lemma 2.6 ζ_t is ψ-irreducible, weak Feller and $supp(\psi)$ has a nonempty interior. Here one can consider the trivial topology define on $\mathcal{K}$, i.e. for every $s \in \mathcal{K}$ has a neighborhood-system $\{s\}$ and all the subsets of $\mathcal{K}$ containing s. Hence for this topology every subset of $\mathcal{K}$ is open and closed and compact too. For $\mathbb{R}^p$ we can consider, e.g. any usual topology, and therefore, derive the product topology for the system.
By Proposition $6.2.8$ $(ii$) all compact subsets of Ω are petite. However, ζ_t is an irreducible and aperiodic chain for which all compact set are petite, it follows from Theorem $5.5.7$ that all compact sets are small set of ζ_t. We have then proved the existence of a small set for the chain. ∎

After the proof of the existence of small set it remain to make a suitable choice of this set and also a choice of the stabilization function to achieve the geometric ergodicity. For this purpose, we proceed in two steps. We first consider the case where $p = 1$ and later on extend the results obtained for $p = 1$ to the case $p \geq 1$.

2.5.3 Geometric Ergodic Conditions for First Order GMAR-ARCH

Let us consider $p = 1$ and define $g\binom{S}{U} = 1 + \|U\|^2$, i.e.,

$$\begin{aligned} g(\zeta_t) &= 1 + X_t^2 \\ &= 1 + \left(\sum_{k=1}^{K} S_{t,k}(m_k(X_{t-1}) + \sigma_k(X_{t-1})\epsilon_t) \right)^2 \\ &= 1 + \sum_{k=1}^{K} S_{t,k} \left(m_k^2(X_{t-1}) + \sigma_k^2(X_{t-1})\epsilon_t^2 + 2m_k(X_{t-1})\sigma_k(X_{t-1}\varepsilon_t) \right). \end{aligned}$$

We obtain the following theorem

Theorem 2.9 *Let $p = 1$ and suppose **A**.2.4 holds. If*

$$\max_{l\in\{1,\cdots,K\}} \limsup_{|x|\to\infty} \frac{\sum_k \left(m_k^2(x) + \sigma_k^2(x)\right) a_{l,k}}{x^2} \quad < \quad 1,$$

with $\mathbb{P}(Q_t = k \mid Q_{t-1} = l) = a_{l,k}$. Then ζ_t is geometrically ergodic.

To a certain extent this theorem is the analog of the Proposition 2.1 in Stockis et al [83] where they considered a mixture of first order NAR-ARCH and found conditions under which the geometric ergodic property is satisfied.

Proof: The existence of a small set is provided by the previous lemma. To conclude the proof, we just need to make a suitable choice of such a set and find a function $g(\zeta) \geq 1$ $\beta > 0$ and a constant $M > 0$ such that

$$\frac{\mathbb{E}\left(g(\zeta_t) \mid \zeta_{t-1} = \binom{s}{x}\right) - g\binom{s}{x}}{g\binom{s}{x}} \leq -\beta \quad \text{for} \quad \|\zeta_{t-1}\| > M.$$

Consider g as defined before the theorem, it follows that

$$\begin{aligned}
& \frac{\mathbb{E}\left(g(\zeta_t) \mid \zeta_{t-1} = \binom{s}{x}\right) - g\binom{s}{x}}{g\binom{s}{x}} \\
= \quad & \frac{\sum_k ((m_k^2(x) + \sigma_k^2(x))\, \mathbb{E}(S_{t,k} \mid S_{t-1} = s) - x^2}{1 + x^2} \\
\leq \quad & \max_l \frac{\sum_k ((m_k^2(x) + \sigma_k^2(x))\, a_{l,k}}{x^2} - 1 \\
\leq \quad & -\beta
\end{aligned}$$

for $\|\zeta\| > M$ as $\max_{l\in\{1,\cdots,K\}} \limsup_{|x|\to\infty} \frac{\sum_k \left(m_k^2(x)+\sigma_k^2(x)\right)a_{l,k}}{x^2} < 1$. ∎

2.5.4 Geometric Ergodic Conditions for Higher Order GMAR-ARCH

Let us now consider the case where $p \geq 1$ and assume the decomposition of the m_k and σ_k as stated in in **A**.2.4. In this context we reformulate the above theorem to derive the following analogous.

Theorem 2.10 *Consider the Markov Chain defined by $\{\zeta_t\}$. If **A**. 2.4 holds and*

$$\max_{l\in\{1,\cdots,K\}} \left\{ \sum_{k=1}^{K} a_{l,k} \left[\left(\sum_{i=1}^{p} |\alpha_{i,k}| \right)^2 + \sum_{i=1}^{p} d_{i,k} \right] \right\} < 1,$$

*where the $\alpha_{i,k}, \beta_{i,k}$ are coefficients considered in the decompositions assumed in **A**. 2.4. Then $\{\zeta_t\}$ is geometric ergodic.*

The theorem can be interpreted in different ways, but we will rather point out the simple fact that the process can remain stationary if the probability of moving from a non stationary state to a stationary one is high enough and conversely if the probability of moving from a stationary state to a non stationary state is small enough. This is much clearer if we consider the simple case of independent mixture, i.e. the case where the processes $S_{t,k}$ are considered as independent sequences of stationary random variables. In this case the process will be stationary if the probability of being in a non stationary state is very small and that of being in a stationary one is high enough. Before we give a proof of the theorem, let us state a corollary related to the latter case.

Corollary 2.11 *Consider the process $\{\zeta_t\}$ with S_t a sequence of i.i.d. random variables, suppose* ***A.*** *2.4 holds and*

$$\sum_{k=1}^{K} \pi_k \left(\left(\sum_{i=1}^{p} |\alpha_{i,k}| \right)^2 + \sum_{i=1}^{p} d_{k,i} \right) < 1,$$

where π_k is the probability of being in the state k. Then $\{\zeta_t\}$ is geometric ergodic.

From this corollary we can derive the following one if we consider a pure mixture of ARCH (p) processes.

Corollary 2.12 *For a mixture of K ARCH (p) processes, i.e. $m_k \equiv 0$ for all k and $\sigma_k^2(u) = \omega_k + \sum_{i=1}^{p} d_{k,i} u_i^2$ and S_t a sequence of i.i.d. random variables, if*

$$\sum_{k=1}^{K} \pi_k \left(\sum_{i=1}^{p} d_{k,i} \right) < 1,$$

then $\{\zeta_t\}$ is geometric ergodic.

These corollaries follow directly from the previous theorem by considering the special case where the the Hidden Markov Chain is considered a sequence of i.i.d. random variables.Therefore, we rather present a detailed proof of Theorem 2.10.

Proof: In the proof of this theorem we will assume the existence of a small set first established in the previous lemma. The rest of the proof consists of finding a suitable function g and a small set as above. Let us recall that

$$U_t = \begin{pmatrix} X_t \\ . \\ . \\ . \\ X_{t+1-p} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_{t,k}) \\ X_{t-1} \\ . \\ . \\ X_{t+1-p} \end{pmatrix},$$

$$\zeta_t = \begin{pmatrix} S_t \\ U_t \end{pmatrix}$$

and define

$$\begin{aligned} g(\zeta_t) &= 1 + h(U_t) \\ &= 1 + X_t^2 + b_{p-1}X_{t-1}^2 + \cdots + b_1 X_{t-p+1}^2, \end{aligned}$$

where the coefficients $b_{p-1}, \cdots, b_1$ will be suitably chosen later. Now, given

$$\max_{l \in \{1,\cdots,K\}} \left\{ \sum_{k=1}^{K} a_{l,k} \left[\left(\sum_{i=1}^{p} |\alpha_{i,k}| \right)^2 + \sum_{i=1}^{p} d_{k,i} \right] \right\} < 1,$$

let us now compute

$$\mathbb{E}\left(g(\zeta_{t+1}) \mid U_t = u, S_t = s\right),$$

i.e.

$$\begin{aligned} \mathbb{E}\left(g(\zeta_{t+1}) \mid U_t = u, S_t = s\right) &= 1 + \mathbb{E}\left(X_{t+1}^2 \mid U_t = u, S_t = s\right) \\ &\quad + b_{p-1}X_t^2 + \cdots + b_1 X_{t-p+2}^2. \end{aligned}$$

To do so, let us first compute

$$\begin{aligned} & \mathbb{E}\left(X_{t+1}^2 \mid U_t = u, S_t = s\right) \\ = \; & \mathbb{E}\left(\sum_{k=1}^{K} S_{t+1,k}(m_k^2(U_t) + \sigma_k^2(U_t)) \mid U_t = u, S_t = s \right) \\ = \; & \sum_{k=1}^{K} a_{l,k}(m_k^2(u) + \sigma_k^2(u)) \end{aligned}$$

by an application of the stationary assumption on S_t, l denotes the non-vanishing coordinate of s. Now using the linear decomposition on the autoregressive and volatility functions defined **A.** 2.4, Equation 4, it follows that

$$\begin{aligned} & \mathbb{E}\left(X_{t+1}^2 \mid U_t = u, S_t = s\right) \\ = \; & \sum_{k=1}^{K} a_{l,k} \left\{ \left(\sum_{i=1}^{p} \alpha_{k,i} u_i + o(\|u\|) \right)^2 + \sum_{i=1}^{p} d_{k,i} u_i^2 + o(\|u\|^2) \right\} \\ = \; & \sum_{i=1}^{p} \left(\sum_{k=1}^{K} a_{l,k}(\alpha_{k,i}^2 + d_{i,k}) \right) u_i^2 + 2 \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} \left(\sum_{k=1}^{K} a_{l,k} \alpha_{i,k} \alpha_{j,k} \right) u_i u_j \\ + \; & o\left(\|u\|\right) \sum_{i=1}^{p} \left(\sum_{k=1}^{K} a_{l,k} \alpha_{i,k} \right) u_i + o\left(\|u\|^2\right) \end{aligned}$$

Since $ab \leq |ab|$ and $2ab \leq a^2 + b^2$ for all a, b real numbers, we can decompose the above expectation in the following way.

$$\begin{aligned}\mathbb{E}\left(X_{t+1}^2 \mid U_t = u, S_t = s\right) &\leq \sum_{i=1}^{p}\left(\sum_{k=1}^{K} a_{l,k}(\alpha_{k,i}^2 + d_{i,k}) + b_i\right) u_i^2 \\ &+ \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\sum_{k=1}^{K} a_{l,k}\left|\alpha_{i,k}\alpha_{j,k}\right|\left(u_i^2 + u_j^2\right) + o\left(\|u\|^2\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{E}\left(g(\zeta_{t+1}) \mid U_t = u, S_t = s\right) \\ \leq\ & 1 + \left(\sum_{k=1}^{K} a_{l,k}(\alpha_{p,k}^2 + d_{p,k} + |\alpha_{p,k}|\sum_{j\neq p}|\alpha_{j,k}|)\right) u_p^2 \\ +\ & \sum_{i=2}^{p-1}\left(\sum_{k=1}^{K} a_{l,k}(\alpha_{k,i}^2 + d_{k,i} + |\alpha_{k,i}|\sum_{j\neq i}|\alpha_{j,k}|) + b_i\right) u_i^2 \\ +\ & \left(\sum_{k=1}^{K} a_{l,k}(\alpha_{1,k}^2 + d_{1,k} + |\alpha_{1,k}|\sum_{j\neq 1}|\alpha_{1,k}|) + b_1\right) u_1^2 + o\left(\|u\|^2\right),\end{aligned}$$

i.e.

$$\begin{aligned}&\mathbb{E}\left(g(\zeta_{t+1}) \mid U_t = u, S_t = s\right) \\ \leq\ & 1 + \left(\frac{\sum_{k=1}^{K} a_{l,k}(\alpha_{p,k}^2 + d_{p,k} + |\alpha_{p,k}|\sum_{j\neq p}|\alpha_{j,k}|)}{b_{p-1}}\right) b_{p-1}u_p^2 \\ +\ & \sum_{i=2}^{p-1}\left(\frac{\sum_{k=1}^{K} a_{l,k}(\alpha_{k,i}^2 + d_{k,i} + |\alpha_{k,i}|\sum_{j\neq i}|\alpha_{j,k}|) + b_i}{b_{i-1}}\right) b_{i-1}u_i^2 \\ +\ & \left(\sum_{k=1}^{K} a_{l,k}(\alpha_{1,k}^2 + d_{1,k} + |\alpha_{1,k}|\sum_{j\neq 1}|\alpha_{1,k}|) + b_1\right) u_1^2 + o\left(\|u\|^2\right).\end{aligned}$$

Choose the b_i in such a way that

$$\max_l \sum_{k=1}^{K} a_{l,k}(\alpha_{1,k}^2 + d_{1,k} + |\alpha_{1,k}|\sum_{j\neq 1}|\alpha_{1,k}|) + b_1 < 1,$$

$$\max_l \frac{\sum_{k=1}^{K} a_{l,k}(\alpha_{k,i}^2 + d_{k,i} + |\alpha_{k,i}|\sum_{j\neq i}|\alpha_{j,k}|) + b_i}{b_{i-1}} < 1 \quad i = 2, 3, \cdots, p-1$$

$$\max_l \frac{\sum_{k=1}^{K} a_{l,k}(\alpha_{p,k}^2 + d_{p,k} + |\alpha_{p,k}|\sum_{j\neq p}|\alpha_{j,k}|)}{b_{p-1}} < 1.$$

Such b_i exits due to the assumption that

$$\max_{l\in\{1,\cdots,K\}}\left\{\sum_{k=1}^{K}a_{l,k}\left[\left(\sum_{i=1}^{p}|\alpha_{i,k}|\right)^2+\sum_{i=1}^{p}d_{k,i}\right]\right\}<1.$$

Moreover, we can choose them in such a way that

$$\max_{l}\sum_{k=1}^{K}a_{l,k}(\alpha_{p,k}^2+d_{p,k}+|\alpha_{p,k}|\sum_{j\neq p}|\alpha_{j,k}|)<b_{p-1}$$

and

$$\begin{aligned}&\max_{l}\sum_{k=1}^{K}a_{l,k}(\alpha_{k,i}^2+d_{k,i}+|\alpha_{k,i}|\sum_{j\neq i}|\alpha_{j,k}|)\\&<b_{i-1}<\\&1-\max_{l}\sum_{v=1}^{i}\sum_{k=1}^{K}a_{l,k}(\alpha_{v,k}^2+d_{v,k}+|\alpha_{v,k}|\sum_{j\neq v}|\alpha_{v,k}|)\end{aligned}$$

for $i=1,\cdots,p-1$. Taking

$$\begin{aligned}r &= max\left\{\max_{l}\sum_{k=1}^{K}a_{l,k}(\alpha_{1,k}^2+d_{1,k}+|\alpha_{1,k}|\sum_{j\neq 1}|\alpha_{1,k}|)+b_1,\right.\\&\quad\max_{l}\frac{\sum_{k=1}^{K}a_{l,k}(\alpha_{k,i}^2+d_{k,i}+|\alpha_{k,i}|\sum_{j\neq i}|\alpha_{j,k}|)+b_i}{b_{i-1}},\\&\quad\text{for}\quad i=2,\cdots,p-1\\&\quad\left.\max_{l}\frac{\sum_{k=1}^{K}a_{l,k}(\alpha_{p,k}^2+d_{p,k}+|\alpha_{p,k}|\sum_{j\neq p}|\alpha_{j,k}|)}{b_{p-1}}\right\}\end{aligned}$$

it follows that

$$\begin{aligned}\mathbb{E}\left(g(\zeta_{t+1})\quad|\quad\zeta_t=\begin{pmatrix}S\\U\end{pmatrix}\right)&\leq rg\begin{pmatrix}S\\U\end{pmatrix}+o\left(\|U\|^2\right)\\&=(r+o(1))g\begin{pmatrix}S\\U\end{pmatrix}-1.\end{aligned}$$

Then, we need to choose δ large enough, such that $r+o(1)<r_0<1\quad$ for $\quad\|\zeta\|>\delta$. Setting $C=\{\|\zeta\|\leq\delta\}$, we obtain the small set with positive measure and conclude the proof. ∎

2.6 Some Applications

Let us consider some examples.
First, we are given as observation a single process, e.g., an AR-ARCH (1), i.e. a process of the type

$$X_t = \alpha X_{t-1} + \sqrt{(\beta + dX_{t-1}^2)}Z_t, t \in \mathbb{Z} \tag{2.14}$$

where the Z_t are i.i.d. distributed with mean zero and unit variance.The geometric ergodic condition for such a model can be summarized by the following condition

$$\alpha^2 + d < 1.$$

This equation represents the classical stationarity condition for a first order AR-ARCH process.
If we, however consider a mixture of two AR-ARCH (1) processes with constant autoregressive coefficients α_1 and α_2, volatility coefficients $\beta_1, \beta_2, d_1, d_2$ and constant state probabilities $0 < \pi < 1$ of being in one state and $1 - \pi$ of being in the other, i.e. we assume

$$X_t = \begin{cases} \alpha_1 X_{t-1} + \sqrt{(\beta_1 + d_1 X_{t-1}^2)}\epsilon_t & \text{with probability} \quad \pi \\ \alpha_2 X_{t-1} + \sqrt{(\beta_2 + d_2 X_{t-1}^2)}\zeta_t & \text{with probability} \quad 1 - \pi \end{cases}$$

with ϵ_t, ζ_t independent sequences of i.i.d. random variables with zero expectations and unit variances. Then, the geometric ergodic condition for this independent mixture is given by the representation

$$\pi(\alpha_1^2 + d_1) + (1 - \pi)(\alpha_2^2 + d_2) < 1.$$

If $\pi \approx 1$, then the parameter of the second AR-ARCH process may violate the stationarity condition, i.e. $\alpha_2^2 + d_2^2 > 1$, though the mixture process is geometric ergodic.
Before we present some other corollaries of the previous theorem, let us define and comment some mixing conditions.

2.6.1 Mixing Conditions

Definition 2.6.1 *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{B}$ and $\mathcal{C}$ two sub sigma-algebra of $\mathcal{A}$. Let us define*

$$\alpha = \alpha(\mathcal{B}, \mathcal{C}) = \sup_{\substack{B \in \mathcal{B} \\ C \in \mathcal{C}}} |\mathbb{P}(B \cap C) - \mathbb{P}(B)\mathbb{P}(C)|$$

*$\{X_t, t \in \mathbb{Z}\}$ is said to be α-**mixing** or (**strongly mixing**) if*

$$\alpha_k = \sup_{t \in \mathbb{Z}} \alpha(\sigma(X_s, s \leq t), \sigma(X_s, s \geq t + k)) \underset{k \to \infty}{\to} 0 \, . \tag{2.15}$$

The process $\{X_t\}$ is said to be α-mixing with geometrically decreasing mixing coefficients if $\alpha_k \leq a_1 e^{-a_2 k}$, $k \geq 1$ for some a_1, $a_2 > 0$.

If $\{X_t, t \in \mathbb{Z}\}$ is stationary, the sup can be omitted. As one can remark, the mixing property for a given series can be considered as an asymptotic measure of independence. More details on this topic can be found in Doukhan [18] and Bosq [10] among others.

Corollary 2.13 *Under the assumption of Theorem 2.10 the process $\{\zeta_t\}$ is exponentially α-mixing. Hence, $\{S_t\}$ and $\{X_t\}$ are also exponentially α-mixing*

Proof: The proof follows by observing that $\{\zeta_t\}$ is geometric ergodic and using the lemma by Davydov ([14], 1973). In this lemma he proved that every geometric ergodic Markov Chain $\{\zeta_t\}$ for which $\{\zeta_0\}$ is distributed according to its initial stationary distribution, is exponentially α-mixing. ∎

In this section we have set conditions that ensure the geometric ergodic property of our model. We have therefore provided conditions for which the model is asymptotic stationary and satisfies some mixing conditions. This result has confirmed our intuition that a process can be considered (globally) as stationary although some phases of the process are not stationary. Furthermore, we can consider some other advantages that we will find in some applications, namely in data analysis or data mining. Usually people use some rules of thumb to remove the outlier and work with the rest of the data, but under our setting we will expect one of the sub-models to get rid of them. However, we need to mention that the choice of the autoregressive order p, the order of the hidden process and its number of state are well-known problems and may be interesting and exciting area for further investigations.

3 Neural Networks and Universal Approximation

In this section we discuss the universal approximation property of some classes of parametric functions and apply the general theory to some classes of Neural Network functions. Let us first present a more general theory and its related results.

3.1 Universal Approximation for some Parametric Classes of Functions

The aim of this section is to study the general problem of estimating the autoregressive functions by fitting functions from parametric classes increasing with the sample size.

3.1.1 Generalities

In this subsection we essentially present and comment the assumptions we need to establish the universal approximation property delivered by some classes of parametric functions. Let us recall the model and the stochastic nature of the hidden process Q_t.

$$X_t = \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_{t,k}) \text{ with } S_{t,k} = \begin{cases} 1 & \text{if } k = Q_t \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Assume the ϵ_t are i.i.d. random variables independent of the past information, ϵ_t and S_t are uncorrelated conditioned on the past information and $\mathbb{E}X_t^2 < \infty$.
Let us define the nonlinear least squares (NLLS) of the m_k as follows

$$\sum_{t=1}^{n}(X_t - \sum_{k=1}^{K} S_{t,k}m_k(\mathbb{X}_{t-1}))^2 \;=\; \sum_{t=1}^{n}\sum_{k=1}^{K} S_{t,k}(X_t - m_k(\mathbb{X}_{t-1}))^2 \tag{3.2}$$

We discuss the problem of estimating at each time instant the autoregressive functions m_k by fitting some functions from parametric classes increasing with sample size n.
Based on the observed process X_t and the hidden process S_t we will consider the stochastic process $\{(X_t, \mathbb{X}_{t-1}, S_t), \quad t \in \mathbb{Z}\}$ defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $X_t \in \mathbb{R}$, $\mathbb{X}_{t-1} \in \mathbb{R}^p$ and $S_t = (S_{t,1}, \cdots, S_{t,K}) \in \mathcal{K}$. It follows by equation 3.1, that at each time instant one and only one component of S_t takes the value 1 and the others take the value 0.
Let $\mathcal{F}_{-\infty}^{s}$ be the σ-algebra generated by (X_t, S_t) for $t \leq s$ and let $\mathcal{F}_{s}^{\infty}$ the σ-algebra generated by (X_t, S_t) for $t \geq s$.

A. 3.1 *Let us assume (X_t, S_t) to be α-mixing with geometrically decreasing rate.*

One can refer to Chapter 2, Definition 2.6.1 to an introduction on the mixing conditions. In fact, in the previous chapter, it was proved that this assumption holds under mild conditions.

Remark 3.2 *Assumption **A**.3.1 implies that the process* $\{(X_t, \mathbb{X}_{t-1}, S_t)\}$ *is* α*-mixing with mixing coefficients, up to a constant factor (i.e.* $\alpha_Y(k) \leq \alpha_\zeta(k-p)$*), decreasing as those of* $\{(X_t, S_t)\}$.

Proof: To see the above remark, let us define $\zeta_t = (X_t, S_t)$ and $Y_t = (X_t, \mathbb{X}_{t-1}, S_t)$. It follows that

$$\mathcal{F}_{-\infty}^{n}(Y) = \sigma(Y_t, t \leq n) \subseteq \sigma(\zeta_t, t \leq n) = \mathcal{F}_{-\infty}^{n}(\zeta)$$

and

$$\mathcal{F}_{n+k}^{\infty}(Y) = \sigma(Y_t, t \geq n+k) \subseteq \sigma(\zeta_t, t \geq n+k-p) = \mathcal{F}_{n+k-p}^{\infty}(\zeta).$$

Furthermore, it follows by definition of the mixing coefficient that $\alpha_Y(k) \leq \alpha_\zeta(k-p)$ for $k-p>0$. ∎

Now, let us consider $\mathcal{G}_{n,k}$ $k = 1, \cdots, K$, to be increasing classes of functions, each containing the null function and defined from $\mathbb{R}^p \longrightarrow \mathbb{R}$, depending on the sample size. Let us also define increasing classes of functions

$$\mathcal{G}_n = \{g = (g_1, \cdots, g_K); g_k \in \mathcal{G}_{n,k}, k = 1, \cdots, K\},$$

$$\mathcal{D}_n = \{sg^T; s \in \mathcal{K}, g \in \mathcal{G}_n\}$$

At each time instant we want to estimate the suitable m_k by minimizing the average nonlinear least squares error, i.e.

$$\begin{aligned} m_n &= (m_{n,1}, \cdots, m_{n,K}) \\ &= \arg\min_{g \in \mathcal{G}_n} \frac{1}{n} \sum_{t=1}^{n} \sum_{k=1}^{K} S_{t,k}(X_t - g_k(\mathbb{X}_{t-1}))^2 \end{aligned} \tag{3.3}$$

If we were to set $K = 1$, we should realize that this problem goes back to the broad class of non-parametric regression estimates based on Grenander's method of Sieves. Now, to claim the consistency of such function estimates we will assume the denseness of $\mathcal{D}_\infty$ in $L^2(\lambda)$, i.e. the space of square integrable functions on $\mathbb{R}^p \times \mathcal{K}$ w.r.t. λ the stationary law of $(\mathbb{X}_t, S_t)$.
If we set $d_n(z,s) = m_n(z)s^T$, then d_n minimizes over all $d = gs^T \in \mathcal{D}_n$

$$\frac{1}{n} \sum_{t=1}^{n} \left(X_t S_t e^T - d(\mathbb{X}_{t-1}, S_t)\right)^2 = \frac{1}{n} \sum_{t=1}^{n} \sum_{k=1}^{K} S_{t,k} \left(X_t - g_k(\mathbb{X}_{t-1})\right)^2 \tag{3.4}$$

where $e = (1, \cdots, 1) \in \mathbb{R}^K$ and where we have used that exactly one $S_{t,k} = 1$ and the others are 0. Using Lemma 10.1 of Györfy et al [39] one can prove that for $d_\infty(z, s) = m(z)s^T$

$$\begin{aligned} & \int\int |d_n(z,s) - d_\infty(z,s)|^2 \lambda(dz, ds) \\ \leq \; & 2 \sup_{d \in \mathcal{D}_n} \left| \frac{1}{n} \sum_{t=1}^{n} \sum_{k=1}^{K} (X_t S_t e^T - d(\mathbb{X}_{t-1}, S_t))^2 - \mathbb{E} \sum_{k=1}^{K} (X_1 S_1 e^T - d(\mathbb{X}_0, S_1))^2 \right| \\ + \; & \inf_{d \in \mathcal{D}_n} \int\int |d_n(z,s) - d_\infty(z,s)|^2 \lambda(dz, ds) \end{aligned} \tag{3.5}$$

i.e. the integrated squared error is bounded by a random estimation error and $L^2(\lambda)$ approximation error.
As example, we can consider the $\mathcal{G}_{n,k}$ to be given sets of feedforward networks, which we will discuss in the next section, or as series expansions

$$\mathcal{G}_{n,k} = \left\{ \sum_{j=1}^{H_{n_k}} a_j \Psi_j, \quad a_1, \cdots, a_{H_{n_k}} \in \mathbb{R}, \quad \sum_{t=1}^{H_{n_k}} |a_j| \leq \triangle_{n_k} \right\}$$

for some given basis $\{\Psi_i\}$ of functions of $L^2(\mu)$, satisfying the denseness assumption on $\mathcal{G}_{k,\infty} = \cup_{n \geq 1} \mathcal{G}_{n,k}$ for $H_{n_k} \longrightarrow \infty, \triangle_{n_k} \longrightarrow \infty$ for all k. White and Wooldridge [92], or recently Györfy et al [39] considered this class of functions without the assumption of the change in the dynamic of the observed process.
Furthermore, if we consider $n_1, \cdots, n_K$ to be the number of realizations for each dynamic of the process, we have

$$n_k = \sum_{k=1}^{n} S_{t,k}$$

and it follows that

$$n_1 + n_2 + \cdots + n_K = n.$$

Since S_t is considered as a stationary Markov chain and by mean of some other considerations on ergodic processes, the strong law of large number or the Ergodic Theorem then implies,

$$\frac{n_k}{n} \underset{a.s}{\longrightarrow} \pi_k \quad \text{as} \quad n \longrightarrow \infty.$$

For the consistency of estimates, instead of assuming uniform boundedness of the function in $\mathcal{G}_\infty$ (as in the universal approximation theory of Neural Networks as presented by White, e.g. in [91]), we will follow the approach of Györfy et al [39]. Before we move into this direction, let us recall some definitions. Nevertheless, for detailed information, one can refer to [39] on the following issues.

3.1.2 Excursion to L_p Norm Covers and VC Dimension

In this section we provide a definition of the norm cover and also of the Vapnik-Chervonenkis (VC) dimension. For this purpose, let us consider $\epsilon > 0$, $\mathcal{G}$ be a set of functions from $\mathbb{R}^d \longrightarrow \mathbb{R}$, $\quad 1 \leq p < \infty$ and ν a probability measure on $\mathbb{R}^d$. Now, we can state the following

Definition 3.1.1 *For a function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ set*

$$\|f\|_{L_p(\nu)} := \left\{ \int |f(z)|^p d\nu \right\}^{\frac{1}{p}}.$$

Then, every collection of function $g_1, \cdots, g_N : \mathbb{R}^d \longrightarrow \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there exists a $j = j(g) \in \{1, \cdots, N\}$ such that

$$\|g - g_j\|_{L_p(\nu)} < \varepsilon$$

is called an ε-cover of $\mathcal{G}$ with respect to the $\| \ \|_{L_p(\nu)}$

Analog definition can be given with respect to $\| \ \|_\infty$. Instead of doing we now consider $Z_1^n = (Z_1, \cdots, Z_n)$ to be n fixed points in $\mathbb{R}^d$. Let ν_n be the corresponding empirical measure, i.e.

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(Z_i) \quad (A \subseteq \mathbb{R}^d)$$

and state the following definitions.

Definition 3.1.2 *a) Considering*

$$\|f\|_{L_p(\nu_n)} = \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(Z_i)| \right\}^{1/p},$$

any ϵ-cover of $\mathcal{G}$ w.r.t. $\|.\|_{L_p(\nu_n)}$ will be called an $L_p\epsilon$-cover of $\mathcal{G}$ on Z_1^n and the ϵ-covering number of $\mathcal{G}$ w.r.t. $\|.\|_{L_p(\nu_n)}$ will be denoted by

$$\mathcal{N}(\epsilon, \mathcal{G}, Z_1^n),$$

i.e. $\mathcal{N}(\epsilon, \mathcal{G}, Z_1^n)$ is the minimal $n \in \mathbb{N}$ such that there exist functions $g_1, \cdots, g_n : \mathbb{R}^d \longrightarrow \mathbb{R}$ with the property that for every $g \in \mathcal{G}$ there is a $j = j(g) \in \{1, \cdots, n\}$ such that

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} |g(Z_i) - g_j(Z_i)|^p \right\}^{1/p} < \epsilon.$$

b) Now, let $\mathcal{A}$ be a class of subsets of $\mathbb{R}^d$ and $n \in \mathbb{N}$. Let $Z_1, \cdots, Z_n \in \mathbb{R}^d$ and define

$$S(\mathcal{A}, \{Z_1, \cdots, Z_n\}) = |\{A \cap \{Z_1, \cdots, Z_n\} \ : \ A \in \mathcal{A}\}|,$$

that is, $S(\mathcal{A}, \{Z_1, \cdots, Z_n\})$ is the number of different subsets of $\{Z_1, \cdots, Z_n\}$ of the form $A \cap \{Z_1, \cdots, Z_n\}$ $A \in \mathcal{A}$.
Then, the nth shatter coefficient of $\mathcal{A}$ is

$$S(\mathcal{A}, n) = \max_{\{Z_1, \cdots, Z_n\} \in \mathbb{R}^d} S(\mathcal{A}, \{Z_1, \cdots, Z_n\}).$$

That is, the shatter coefficient is the maximal number of different subsets of n points that can be picked out by sets from $\mathcal{A}$.

c) Finally, if we assume $\mathcal{A} \neq \emptyset$, the VC dimension of $\mathcal{A}$ or $V_{\mathcal{A}}$ is defined by

$$V_{\mathcal{A}} = \sup\{n \in \mathbb{N} : S(\mathcal{A}, n) = 2^n\}.$$

Thus, the VC dimension $V_{\mathcal{A}}$ is the largest integer n such that there exists a set G of n points in $\mathbb{R}^d$ that can be shattered by $\mathcal{A}$, i.e.

$$S(\mathcal{A}, G) = 2^n.$$

In the remainder we will consider the L_1 norm and use the definitions as stated in this section.

3.1.3 Consistency of Least Squares Estimates

In this section we follow the approach by Györfy et al [39] where the original least squares estimate m_n is replaced by a truncated version, that is for some sequence $\triangle_n \longrightarrow \infty$ we consider

$$\hat{m}_n(z) = T_{\triangle_n} m_n(z) = (T_{\triangle_n} m_{n,1}(z), \cdots, T_{\triangle_n} m_{n,K}(z)), \tag{3.6}$$

where the truncation operator T_L is defined as

$$T_L(y) = \begin{cases} y & \text{if} \quad |y| \leq L, \\ L \times sign(y) & \text{otherwise.} \end{cases}$$

We follow this approach as used in Franke et al [29] and extend it to the case of GMAR-ARCH models, models like those defined in Equation 3.1. For this purpose let us define

$$\hat{\mathcal{G}}_n = \{T_{\triangle_n} g(z), \quad g \in \mathcal{G}_n\}$$

as the class of truncated functions of $\mathcal{G}_n$. and correspondingly

$$\hat{\mathcal{D}}_n = \{d = sg^T; s \in \mathcal{K}, g \in \hat{\mathcal{G}}_n\}.$$

We will assume the following

A. 3.3 *$\hat{\mathcal{D}}_n$ is a class of bounded real-valued functions on $\mathbb{R}^p \times \mathcal{K}$ such that for all $\delta > 0, n \geq 1$ there exists $K_n(\delta)$ such that for all $z_1, \cdots, z_n \in \mathbb{R}^p, s_1, \cdots, s_n \in \mathcal{K}$ there are $d_l^* \in \hat{\mathcal{D}}_{n,l}, l = 1, \cdots, K_n(\delta)$ with*

$$\forall\ d \in \hat{\mathcal{D}}_n \text{ there is } K_n(\delta) \text{ such that } \frac{1}{n}\sum_{j=1}^{n} |d(z_j, s_j) - d_l^*(z_j, s_j)| < \delta \qquad (3.7)$$

$K_n(\delta)$ is a bound on the $L_1\delta$-covering number w.r.t. the empirical measure of $(z_j, s_j), j = 1 \cdots, n$, holding uniformly in that points. we can derive such a bound from the corresponding bounds for the single function classes

$$\hat{\mathcal{G}}_{n,k} = \{T_{\triangle_n} g_k, g_k \in \mathcal{G}_{n,k}\}, k = 1, \cdots, K.$$

Let for all $k \leq K, \delta > 0, n \geq K$ exists $K_{n,k}(\delta/K)$ such that for all $z_1, \cdots, z_n \in \mathbb{R}^p$ there are $f_{k,l}^* \in \hat{\mathcal{G}}_{n,k}, l = 1, \cdots, K_{n,k}(\delta/K)$ with:
for any $f_k \in \hat{\mathcal{G}}_{n,k}$ there is an $l \leq K_{n,k}(\delta/K)$ such that $\frac{1}{n}\sum_{t=1}^{n} \left|f_k(z_t) - f_{k,l}^*(z_t)\right| < \delta/K$.
Then, it follows that

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n} |d(z_j, s_j) - d_l^*(z_j, s_j)| &= \frac{1}{n}\sum_{j=1}^{n} |\sum_{k=1}^{K} s_{j,k}(f_k(z_j) - f_{k,l_k}^*(z_j)| \\ &\leq \sum_{k=1}^{K} \frac{1}{n}\sum_{j=1}^{n} |(f_k(z_j) - f_{k,l_k}^*(z_j)| < \delta \qquad (3.8) \end{aligned}$$

if we choose $d_l^*(z, s) = \sum_{k=1}^{K} s_k f_{k,l_k}^*(z)$ for suitable $f_{k,l_k}^* \in \hat{\mathcal{G}}_{n,k}$.
We conclude

$$K_n(\delta) \leq \prod_{k=1}^{K} K_{n,k}(\delta/K),$$

compare also Lemma 16.14 of Györfi et al [39] for a similar result.

For later reference as we shall need for the proof of universal approximation, we note (as proved in [29]) that each δ-covering of $\mathcal{D}$ w.r.t. $z_1, \cdots, z_{2n}$ is automatically a 2δ-covering w.r.t. $z_1, \cdots, z_n$, i.e.

$$K_n(2\delta) \leq K_{2n}(\delta), \quad \text{for all} \quad n \geq K, \delta > 0.$$

3.1.4 Universal Approximation

In this section we present the universal approximation property for various classes of parametric functions under the main assumption that the observed process is controlled by a hidden process that has a given fixed number K of states. To state the main result, we need to provide some intermediate and technical results. For this

purpose we extend Theorem 5.1 and Theorem 5.2 of Franke et al [29] to allow the observed process to contain some heterogeneous phases or dynamics. Let us first present a special case of Theorem 5.1 in [29].

Lemma 3.4 *Let (X_t, S_t) be a stationary time series satisfying* **A.** *3.1 and let $Z_t = (X_t, \mathbb{X}_{t-1}, S_t)$ be the corresponding stationary an α-mixing process in $\mathbb{P}^{p+1} \times \mathcal{K}$. Let*

$$\mathcal{G} = \{g = (g_1, \cdots, g_K), g_k \in \mathcal{G}_k, k = 1, \cdots, K\}$$

be a set of measurable functions g with $g_k : \mathbb{P}^{p+1} \to [0, B]$ for some $B > 0$ and let

$$\mathcal{H} = \{h(y, s) = g(y)s^T, y \in \mathbb{P}^{p+1}, s \in \mathcal{K}, g \in \mathcal{G}\}$$

be the corresponding set of real-valued functions on $\mathbb{P}^{p+1} \times \mathcal{K}$. Then, for every $\epsilon > 0, n > 1$

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{t=1}^{n} h(Z_t) - \mathbb{E}h(Z_1)\right| > \epsilon\right) \leq K_{2n}^{H}\left(\frac{\epsilon}{32}\right) c_1 e^{-c_2\sqrt{n}\epsilon/B}$$

where $c_1, c_2 > 0$ are constants not depending on n, and K_{2n}^{H} denotes a bound on the covering number for $\mathcal{H}$.

The above lemma follows from Theorem 5.1 of [29] using that $s_k = 1$ for exactly one and only one k and, therefore, $h : \mathbb{P}^{p+1} \times \mathcal{K} \to [0, B]$, and, by the discussion following **A.** 3.3, $\mathcal{H}$ satisfies **A.** 3.3 too. In particular we have

$$K_n^H \leq \prod_{k=1}^{K} K_{n,k}\left(\delta/K\right)$$

where $K_{n,k}(\delta)$ is a bound on the covering number of $\mathcal{G}_k$ which compose

$$\mathcal{G} = \{g = (g_1, \cdots, g_K), g_k \in \mathcal{G}_k\}.$$

The next result is the analogous of Theorem 5.2 in [29] but under the main assumption of the change in the dynamic of the observed process. As for the latter theorem, this Lemma goes back to Theorem 10.2 of Györfi et al [39] that was established for i.i.d. random variables and extended to time series by Franke et al [29]. Additionally the denseness assumption helps us to get rid of the assumptions 10.9 resp. (10.11) of Theorem 10.2 in [39]. The related lengthy proof follows under our setting without any major changes. We simply use the Ergodic Theorem or the Strong Law of Large numbers for time series, instead of the classical strong law of large numbers. Therefore, we skip the proof here to simplify the presentation.

Lemma 3.5 *Let $\{(X_t, S_t)\}$ be a stationary stochastic process with $X_t \in \mathbb{R}$, $S_t \in \mathcal{K}$ as defined in equation 3.1. Let us consider λ as the stationary distribution of*

$(X_{t-1}, \cdots, X_{t-p}, S_t)$. Let $\mathcal{G}_n \subseteq L^2(\lambda)$ be an increasing class of functions $f = (f_1, \cdots, f_K) : \mathbb{R}^{p+1} \longrightarrow \mathbb{R}^K$, and m_n the corresponding least squares estimate of the autoregressive functions given by equation 3.3. For some sequence of bounds $\triangle_n > 0$ with $\lim_{n\longrightarrow\infty} \triangle_n = \infty$, let $\hat{m}_n = T_{\triangle_n} m_n(z)$ be the truncated least squares estimate of equation 3.6 and let $\hat{\mathcal{G}}_n$ be the set of truncated functions $T_{\triangle_n} f, f \in \mathcal{G}_n$. Assume that the union $\mathcal{G}_\infty$ of $\mathcal{G}_n$ is dense in $L^2(\lambda)$. Define

$$V_t = \sum_{k=1}^{K} S_{t,k}(T_L X_t - f_k(\mathbb{X}_{t-1}))^2$$

with $T_L X_t$, denoting the random variable X_t truncated at $\pm L$.

1. *If for all $L > 0$*

$$\lim_{n\longrightarrow\infty} \mathbb{E}\left\{ \sup_{f\in\hat{\mathcal{G}}_n} \left| \frac{1}{n}\sum_{t=1}^{n} V_t - \mathbb{E}V_1 \right| \right\} = 0$$

then

$$\mathbb{E}\int\int \left(\sum_{k=1}^{K} s_k(m_{n,k}(z) - m_k(z)) \right) \lambda(dz, ds) \longrightarrow 0 \quad (n \longrightarrow \infty).$$

2. *If, additionally, $\{(X_t, S_t)\}$ is ergodic, and if for all $L > 0$*

$$\lim_{n\longrightarrow\infty} \sup_{f\in\hat{\mathcal{G}}_n} \left| \frac{1}{n}\sum_{t=1}^{n} V_t - \mathbb{E}\sum_{k=1}^{K} V_1 \right| = 0 \text{ a.s.},$$

then

$$\lim_{n\longrightarrow\infty} \int\int \left(\sum_{k=1}^{K} s_k(m_{n,k}(z) - m_k(z))^2 \right) \lambda(dz, ds) = 0 \text{ a.s.}.$$

Given the two previous lemma we are in position to state and prove the general result on the universal approximation property of some classes of parametric functions under mild conditions.

Theorem 3.6 *(Universal Approximation)*
*Let (X_t, S_t) as defined in equation 3.1 stationary stochastic process satisfying **A.3.1**, i.e, an α-mixing condition with geometrically decreasing rate. Additionally, assume S_t is an irreducible and aperiodic first order Markov Chain. Let $\mathcal{G}_n, n \geq 1$ be classes of functions in $L^2(\lambda)$, such that their union $\mathcal{G}_\infty$ is dense in $L^2(\lambda)$, and, for*

$\triangle_n \longrightarrow \infty$, the corresponding classes of truncated functions $\hat{\mathcal{G}}_n$ satisfy **A.***3.3. Now, let us define*

$$\kappa_n(\epsilon, n) = K \log K_{2n,1}\left(\frac{\epsilon}{128\triangle_n}\right)$$

where $\hat{\mathcal{G}}_n = \{g = (g_1, \cdots, g_K), g_k \in \hat{\mathcal{G}}_{n,k}\}$ and $K_{2n,1}$ denotes a bound on the covering number of $\hat{\mathcal{G}}_{n,k}, k = 1, \cdots, K$. Let

$$\hat{m}_n(z) = T_{\triangle_n} m_n(z) = (T_{\triangle_n} m_{n,1}, \cdots, T_{\triangle_n} m_{n,K}),$$

i.e. the truncated least squares estimate obtained from equations 3.3 and 3.6.

1. *If, for $n \longrightarrow \infty$, $\triangle_n^2 \kappa_n(\epsilon, n)/\sqrt{n} \longrightarrow 0$ for all $\epsilon > 0$, then*

$$\mathbb{E} \int\int \left(\sum_{k=1}^{K} s_k (\hat{m}_{n,k}(z) - m_k(z))^2\right) \lambda(dz, ds) \longrightarrow 0 \quad (n \longrightarrow \infty).$$

2. *If additionally, $\triangle_n^4 / n^{1-\delta} \longrightarrow 0$ for some $\delta > 0$, then*

$$\int\int \left(\sum_{k=1}^{K} s_k (\hat{m}_{n,k}(z) - m_k(z))^2\right) \lambda(dz, ds) \longrightarrow 0 \quad \textit{a.s.} \quad (n \longrightarrow \infty),$$

   i.e. the approximate is strongly universally consistent.

From this theorem one can say that the consistency proof of some classes of parametric functions can be reduced to the search for bounds on their covering number (assuming the denseness property can be proved for these classes of functions). That is what we shall investigate in the next section for some classes of Neural Networks.

Proof: The proof of this theorem is similar to that of Theorem 2.1 in [29] with Lemma 3.4 and 3.5 replacing Theorem 5.1 and 5.2 in [29], with the slight difference that in this context we assume possible changes in the structure of the observed process and allow the existence of non-stationary phases for this process. Therefore, we have the following probability bound for large sample size with V_t as in Lemma 3.5

$$\mathbb{P}\left(\sup_{f \in \hat{\mathcal{G}}_n} \left| \frac{1}{n} \sum_{t=1}^{n} V_t - \mathbb{E} V_1 \right| > \epsilon \right) \leq K_{2n}^H \left(\frac{\epsilon}{32}\right) c_1 e^{-c_2 n^{1/2} \epsilon / (4\triangle_n^2)},$$

where K_{2n}^H denotes the covering number (compare assumption **A.**3.3) of

$$\mathcal{H}_n = \{h : \mathbb{R}^{p+1} \times \mathcal{K} \longrightarrow \mathbb{R}, h(y, z, s) = \sum_{k=1}^{K} s_k (T_L y - f_k(z))^2 \quad \text{for some} \quad f \in \hat{\mathcal{G}}_n\}.$$

As in the proof Theorem 2.1 in [29], we have

$$K_{2n}^{H}\left(\frac{\epsilon}{32}\right) \leq K_{2n}\left(\frac{\epsilon}{32(4\triangle_n)}\right) = K_{2n}\left(\frac{\epsilon}{128\triangle_n}\right),$$

and recalling that

$$K_{2n}\left(\frac{\epsilon}{128\triangle_n}\right) \leq \prod_{k=1}^{K} K_{2n,k}\left(\frac{\epsilon}{128K\triangle_n}\right) \leq \left(K_{2n,1}\left(\frac{\epsilon}{128K\triangle_n}\right)\right)^{K} = e^{\kappa_n(\epsilon,\triangle_n)}$$

(like in Theorem 2.1 mentioned above) it follows that

$$\mathbb{E}\int\int\left(\sum_{k=1}^{K} s_k(\hat{m}_{n,k}(z) - m_k(z))^2\right)\lambda(dz,ds) \longrightarrow 0 \quad (n \longrightarrow \infty)$$

if $\triangle_n^2\kappa_n(\epsilon,n)/n^{1/2} \longrightarrow 0$.
The proof of the second part of this theorem can be written word for word like that of the strong consistency as presented by Franke et al [29]. It can therefore be omitted here. ∎

3.2 Neural Networks as Universal Approximators

The aim of this section is the application of the universal approximation theory developed in the previous section to some classes of Neural Networks and the derivation of their universal approximation property. For this purpose, let us first focus on the denseness property of some classes of network functions.

3.2.1 Density of Network Classes of Functions

In this section we recall the mathematical definition of the Neural Network, build up special classes of networks to solve our problem and prove its denseness property under mild assumptions. Thus, we want to prove that in our setting, at each time instant, each autoregressive function m_k can be well approximated by a given network function. Let us recall that a network can be defined as follows

$$f_{H_k}(z,\omega_k) = \beta_{o,k} + \sum_{h=1}^{H_k}\beta_{h,k}\psi(z'\gamma_{h,k} + \gamma_{h0,k}),$$

where $\omega_k \in \mathbb{R}^{M(H_k)}$ represents all the weights of the network function with $M(H_k) = H_k(2+p)+1$, and H_k a given number of hidden neurons for this network.
From here on we will consider $S_{t,k}$ as defined in equation 3.1. Let us then define,

$$\mathcal{G}_k(H_k) = \{f_{H_k}(z,\omega_k);\ \omega_k \in \mathbb{R}^{M(H_k)}\},$$

$$\mathcal{G}_k = \{f_{H_k}(z, \omega_k);\ \omega_k \in \mathbb{R}^{M(H_k)},\ H_k \geq 1\},$$

$$\mathcal{G}(H) = \{f = (f_{H_1}, \cdots, f_{H_K});\ f_k \in \mathcal{G}_k(H_k), k = 1, \cdots, K\}$$

where $H = (H_1, \cdots, H_K)$,

$$\mathcal{G} = \{(f_{H_1}, \cdots, f_{H_K});\ f_{H_k} \in \mathcal{G}_k,\ k = 1, \cdots, K\},$$

$$\mathcal{D}(H) = \{f(z, s) = s^T g(z), g \in \mathcal{G}(H)\}$$

and

$$\mathcal{D} = \{f(z, s) = s^T g(z),\ g \in \mathcal{G}\}.$$

In the following we consider only the sigmoid activation functions satisfying

A. 3.7 *ψ is continuous and strictly increasing,*

$$0 < \lim_{x \longrightarrow \infty} \psi(x) = \psi(\infty) \leq 1$$

and

$$-1 \leq \lim_{x \longrightarrow -\infty} \psi(x) = \psi(-\infty) \leq 0.$$

We also define

$$\mathcal{G}_{n,k} = \mathcal{G}_k(H_{n,k})$$

for some increasing sequence $H_{n,k}, n \geq 1$. Obviously $\mathcal{G}_{n,k}$ is increasing with n and their infinite union

$$\mathcal{G}_k = \cup_{n \geq 1} \mathcal{G}_{n,k}$$

satisfies a density property that one can easily prove by a direct application of the Lemma 3.1 in [29]. If we also define correspondingly

$$\mathcal{D}_n = \mathcal{D}(H_n),\ H_n = (H_{n,1}, \cdots, H_{n,K}),$$

it is clearly an increasing sequence and

$$\mathcal{D} \ = \ \cup_{n \geq 1} \mathcal{D}_n \tag{3.9}$$

To apply Theorem 3.6 we need to prove that the functions in $\mathcal{D}$ are universal approximators for a large class of functions in the mean square sense w.r.t. λ. By an extension of Lemma 3.1 in [29] to models with changes in their dynamics, we obtain the following lemma that ensures the denseness property of some classes of network functions.

Lemma 3.8 *(Density of Network Functions)*
Let λ be a measure on $\mathbb{R}^p \times \mathcal{K}$ as defined previously. Let

$$\|f\|_{2,\lambda} = \left\{ \int \int \left(\sum_{k=1}^{K} s_k f_k(z) \right)^2 \lambda(dz, ds) \right\}^{1/2}$$

denote the $L^2(\lambda)$-norm. Let $\mathcal{D}$ be defined by 3.9 with activation function satisfying A.3.7. Then, for any $g \in L^2(\lambda)$ and any $\epsilon > 0$, there exists a function $f \in \mathcal{D}$ such that

$$\|f - g\|_{2,\lambda} < \epsilon$$

Proof: To prove this we can simply observe that, as $|s_k| \leq 1$,

$$\|f - g\|_{2,\lambda} \leq \sum_{k=1}^{K} \|f_k - g_k\|_{2,\mu}$$

where $\| \ \|_{2,\mu}$ represents the $\| \ \|_2$ with respect to the marginal measure μ on $\mathbb{R}^p$ induced by λ. Moreover, for each of the g_k, it follows by Lemma 3.1 in [29] the existence of $f_k \in \mathcal{G}_k$ such that

$$\|f_k - g_k\|_{2,\mu} < \epsilon/K$$

and the proof of the lemma follows. ∎

Once we have stated and proved the previous lemma we have almost all the ingredients needed for the application of Theorem 3.6 to Neural Network classes of functions.

3.2.2 Consistency of Neural Network Estimates

We now consider the nonlinear weighted least squares estimate

$$m_n(z) = (m_{n,1} \cdots, m_{n,K})$$

of $m(z) = (m_1 \cdots, m_K)$ based on feedfoward networks, i.e.

$$m_n(z) = (f_{H_{n,1}}(z; \hat{\omega}_{n,1}) \cdots, f_{H_{n,K}}(z; \hat{\omega}_{n,K})), \tag{3.10}$$

where

$$(\hat{\omega}_{n,1}, \cdots, \hat{\omega}_{n,K}) = \underset{(\omega_1, \cdots, \omega_K) \in \mathbb{R}^{M(H_n)}}{\arg\min} \frac{1}{n} \sum_{t=1}^{n} \sum_{k=1}^{K} S_{t,k} (X_t - f_{H_{n,k}}(\mathbb{X}_{t-1}, \omega_k))^2 \tag{3.11}$$

with $M(H_n) = M(H_{n,1}) + \cdots + M(H_{n,K})$.

Theorem 3.9 *Let $\{(X_t, S_t)\}$ be a stationary stochastic process satisfying* **A.***3.1. For $H_{n,1}, \cdots, H_{n,K} \longrightarrow \infty, \triangle_n \longrightarrow \infty$, let $\hat{m}_n = T_{\triangle_n} m_n$ be the truncated estimate of m given by equations 3.10 and 3.11. Assume ψ satisfies* **A.***3.7, and let $H_n = \min(H_{n,1}, \cdots, H_{n,K})$.*

1. *If for $n \longrightarrow \infty, \triangle_n^2 H_n \log(\triangle_n^2 H_n)/n^{1/2} \longrightarrow 0$ for all $\epsilon > 0$, then*

$$\mathbb{E} \int \int \left(\sum_{k=1}^{K} s_k (m_{n,k}(z) - m_k(z))^2 \right) \lambda(dz, ds) \longrightarrow 0 \quad (n \longrightarrow \infty).$$

2. *If additionally, $\triangle_n^4 / n^{1-\delta} \longrightarrow 0$ for some $\delta > 0$, then*

$$\int \int \left(\sum_{k=1}^{K} s_k (m_{n,k}(z) - m_k(z))^2 \right) \lambda(dz, ds) \longrightarrow 0 \quad \textit{a.s.} \quad (n \longrightarrow \infty).$$

Proof: This proof is partly based on Lemma 3.8 which allows us to prove that $\mathcal{D}$ is dense in $L^2(\lambda)$. Additional, by considering the proof of Theorem 3.6, we only need to find a bound on the covering number of one of the subclasses of the network functions we have considered. For this purpose we make use of 16.19 from the proof of Theorem 16.1 of Györfy et al [39] and the fact that $\hat{\mathcal{G}}_{n,k}$ satisfies **A.**3.3 with

$$K_{2n,k}\left(\frac{\epsilon}{128 \triangle_n}\right) = \left(\frac{12e \triangle_n (H_{n,k}+1)}{\epsilon} 128 \triangle_n\right)^{(2p+5)H_{n,k}+1}$$

w.l.o.g. let us assume that $K_{2n,1}$ is the largest of the $K_{2n,k}, k = 1, \cdots, K$. Regarding the remark after Lemma 3.4, we get as a bound on the covering number of $\hat{\mathcal{D}}$

$$\left(K_{n,1}\left(\frac{\epsilon}{128 \triangle_n K}\right)\right)^K = e^{\kappa_n(\epsilon, \triangle_n)}$$

with

$$\kappa_n(\epsilon, \triangle_n) = K\{(2p+5)H_{n,1}+1\} \log\left(1536K \triangle_n^2 H_{n,1}/\epsilon\right).$$

By neglecting constant factors and terms of smaller order, the rest of the proof follows directly from Theorem 3.6. ∎

In this chapter, under the assumption of change in the dynamic of the observed process, the universal approximation property (of the autoregressive functions by some functions) have been established for some classes of parametric function and in particular for some classes of feedforward networks.However, the S_t are not observed, and therefore, we need a numerical procedure for calculating the $m_{n,k},\ k = 1, \cdots, K$, which is valid for $n \to \infty$. In the next chapter, considering the hidden process S_t in a more general context (than the nonlinear least squares) we will propose a version of the EM algorithm that helps to solve that problem.

4 Hidden Markov Chain Driven Models for Change-point Analysis in Financial Time Series

It is of great importance in mathematics to find conditions that on the one hand are strong enough to have useful consequences, but on the other hand are weak enough to hold (and be easy to check) for many interesting examples. In this spirit, we assume in this chapter that the number of dynamics of the process is known and we also assume that at each time instant given neural networks are able to properly estimate the autoregressive and the volatility functions. We can justify this in certain extent by the universal approximation property established earlier.
In this Chapter, instead of computing the conditional $\log$-likelihood directly, we propose an alternative approach. Indeed, we define a conditional likelihood based on the hidden process and derive an analytical representation under normality assumption of the residuals. For this analytical formulation we explore the asymptotic properties of the parameter estimates given that the autoregressive and volatility function as well can be suitably approximated by given Feedforward Networks. Moreover we present a version of the EM-algorithm that helps us to solve the problem numerically
Let us first recall some definitions and properties related to a discrete first order stationary Markov Chain.

4.1 Discrete Markov Processes

Before we present an intuitive description of this type of process, let us first illustrate its behavior for a two states M.C.

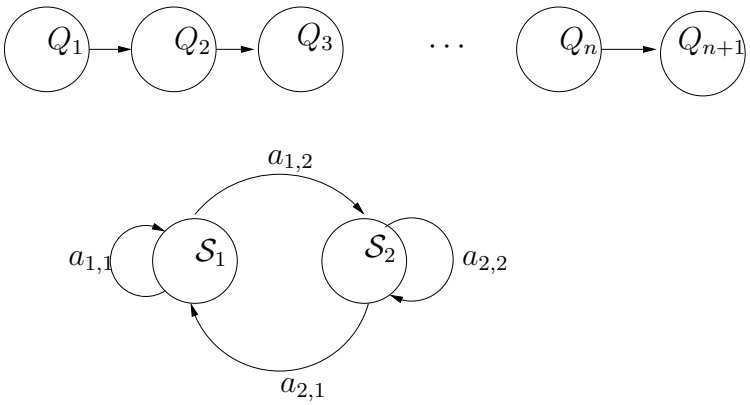

Figure 4.1: Markov Chain

The Markov property is a simple (mathematically tractable) relaxation of the independence assumption. Therefore, for some time series models, it is natural to consider discrete Markov Chains with finite number of states. This type of M.C. can

be described (as illustrated in Figure 4.1) as a system that is at any time instant t in one of its K different states $Q_t = k, k \in I_K$.
At regularly spaced time intervals, a change of state (with the possibility of remaining in the same state), i.e. moving from $Q_t = i$ to $Q_{t+1} = j$ occurs in the system with respect to a set of probabilities associated with the states ($\{a_{i,j}, i, j \in I_K\}$).
In our context we will confine our attention to stationary and homogeneous Markov Chains.
After this somehow intuitive description of M.C., we can present state some useful definitions.

Definition 4.1.1 *Let us consider a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, Let* $I_K = \{1, \cdots, K\}$ *be a finite set,* $(Q_n : n \in \mathbb{N})$ *a collection of random variables taking values in* K*,* $A = (a_{ij} : i, j \in I_K)$ *a stochastic matrix and* π *be a distribution. We call* $(Q_n)_{n \geq 0}$ *a first order Markov Chain with initial distribution* π *and transition matrix* A *if:*

1. Q_0 *has the distribution* π*,*
2. *for* $n \geq 0$*,*

$$\mathbb{P}(Q_{n+1} = j \mid Q_0 = i_0, \ldots, Q_n = i) = \mathbb{P}(Q_{n+1} = j \mid Q_n = i) = a_{ij},$$

with

$$a_{ij} \geq 0, \quad \sum_{j=1}^{K} a_{ij} = 1.$$

That is, the next value of the chain depends only on the current value, not on any previous values. This is often summed up in the pithy phrase, "Markov Chains are memoryless." Still Markov Chain are of great importance in mathematics and many other scientific fields.

Definition 4.1.2 *A Markov Chain* $\{Q_t\}$ *with transition probability matrix* P *is said to be irreducible if* $\forall\, t, \forall\, i, j \in I_K, \exists\, m$ *such that* $P(Q_{t+m} = j \mid Q_t = i) > 0$.

Definition 4.1.3 *A Markov Chain* $\{Q_t\}$ *with transition probability matrix* P *is said to be aperiodic if all of its states are aperiodic, i.e. the period* $d_i = 1\ \forall\ i \in I_K$*, where* $d_i = gcd\{n \geq 1 : (A^n)_{ii} > 0\}$

Definition 4.1.4 *Let* $\{Q_t\}$ *be a Markov Chain with K states and transition probability matrix* P*. A row vector* $\pi = (\pi_1, \cdots, \pi_K)$ *is said to be a stationary distribution for a the Markov Chain, if it satisfies:*

i) $\pi_i \geq 0$ *for all* $i = 1, \cdots, K$ *and* $\sum_{i=1}^{K} \pi_i = 1$*, and*

ii) $\pi P = \pi$*, i.e.* $\sum_{i=1}^{K} \pi_i P_{ij} = \pi_j$ *for* $j = 1, \cdots, K$.

Aperiodicity and irreducibility ensure the existence and uniqueness of such stationary distributions under our setting.

4.2 Hidden Markov Driven Models

The model

$$X_t = \sum_{k=1}^{K} S_{t,k}(\ m_k(\ \mathbb{X}_{t-1}\ ) + \sigma_k(\ \mathbb{X}_{t-1}\ )Z_{t,k}\ ) \text{ with } S_{t,k} = \begin{cases} 1 & \text{for } k = Q_t \\ 0 & \text{otherwise,} \end{cases} \quad (4.1)$$

where K is the given number of states, p the common order of the underlying $NLAR_ARCH$ processes $X_{t,k} = m_k(\ \mathbb{X}_{t-1}\ ) + \sigma_k(\ \mathbb{X}_{t-1}\ )Z_{t,k}$, $m_k : \mathbb{R}^p \to \mathbb{R}$ are the autoregressive functions, $\sigma_k : \mathbb{R}^p \to (\ 0, \infty\ )$ the volatility functions, $\{Z_{t,k}\}$ are i.i.d. random variables with mean zero and variance one , and S_t is a Markov Chain as considered in section 2.

4.2.1 Preliminary Notations

1. The observations $X^n = \{X_t : t = -p+1, \cdots, 1, \cdots, n\}$ are from the time series data collection, n is the number of observations, and t represents each time instant. Similarly we define $\mathbb{X}_{t-1} = \{(\ X_{t-1}, \cdots, X_{t-p}\ )\}$ which is at each time instant the input of the emission model, since our task focuses on autoregressive processes of order p.

2. The number of states K is known. In general, a careful analysis of the model usually provides interpretation for the states in terms of physical significance or economical meaning such as relation to market sentiment, growth, recession, interest rate, volatility, etc.

3. The transition probability matrix

$$A = \{a_{ij},\ i, j \leq K, a_{ij} = \mathbb{P}(\ Q_{t+1} = j \quad | \quad Q_t = i\ )\}, \quad (4.2)$$

   where $a_{ij} \geq 0$, $\sum_j a_{ij} = 1$ and Q_t describes the state at time t.

4. The emission distributions

$$B = \{b_j^t = \mathbb{P}(\ X_t \,|\, \mathbb{X}_{t-1}, Q_t = j\ ),\ 1 \leq j \leq K,\ 1 \leq t \leq n\}. \quad (4.3)$$

5. The initial state probability distribution $\Pi = \{\pi_j, j = 1 \cdots, K\}$ where $\sum_{j=1}^{K} \pi_j = 1$.

6. For each $t = 1, \cdots, n$ we define

$$Q^t = \{Q_s : s = 1, \cdots, t\}$$

   and assume Q_t to be a first order Markov Chain independent of $\mathcal{F}_{t-1} = \sigma\{X_s, s \leq t-1\}$ in the sense that

$$\mathbb{P}(\ X_t \,|\, X^{t-1}, Q^t\ ) = \mathbb{P}(\ X_t \,|\, \mathbb{X}_{t-1}, Q_t\ ) \quad (4.4)$$

and

$$\mathbb{P}(\ Q_t \,|\, Q^{t-1}, X^{t-1}\ ) = \mathbb{P}(\ Q_t \,|\, Q_{t-1}\ ) \tag{4.5}$$

are satisfied.

If we were to assume K, A, B and Π known; one would have been able to use hidden Markov models to generate a full sequence of observations X^n of the process and the procedure could have been described by the following algorithm

Algorithm 4.1 *1) Choose an initial state Q_1 according to the initial state distribution Π, i.e.*

$$\pi_i = P(\ Q_1 = i\ ).$$

2) Set t=1.

3) Generate X_t according to the probability distribution in state i, i.e. $b_i(\ t)$.

4) Transit to a new state $Q_{t+1} = j$ according to the state transition probability for state i, i.e. a_{ij}.

5) Set t=t+1; return to step 3) if $t < n$; otherwise terminate the procedure.

Unfortunately this is not representative for solve the problem in reality. Indeed, we are not able to observe the hidden process, and therefore not able to say exactly where to start, when to make a change and where we should end. Nevertheless, we assume the existence of such a chain and define a version of the conditional likelihood for which we shall find the asymptotic properties of its parameter estimates in our setting.

4.3 Conditional Likelihood

Here we shall take into account the Markov structure of the hidden process and we shall also assume $Q_1, X_0, \cdots, X_{-p+1}$ are given. By doing so let us define the conditional likelihood as follows

$$\begin{aligned} L(\ X \mid Q\ ) &= \mathbb{P}(\ X^n \mid Q^n\ ) && (4.6)\\ &= \mathbb{P}(\ X_n, \cdots, X_1, \cdots, X_{-p+1} \mid Q_n, \cdots, Q_1\ ) && (4.7)\end{aligned}$$

and using an iterative computation it follows that

$$\begin{aligned} L(\ X \mid Q\ ) &= \mathbb{P}(\ X_n \mid X^{n-1}, Q^n\ )\mathbb{P}(\ X^{n-1} \mid Q^n\ )\\ &= \mathbb{P}(\ X_n \mid X^{n-1}, Q^n\ )\frac{\mathbb{P}(\ X^{n-1}, Q^n\ )}{\mathbb{P}(\ Q^n\ )}\\ &= \mathbb{P}(\ X_n \mid X^{n-1}, Q^n\ )\frac{\mathbb{P}(\ Q_n \mid X^{n-1}, Q^{n-1}\ )}{\mathbb{P}(\ Q^n\ )}\mathbb{P}(\ X^{n-1}, Q^{n-1}\ ).\end{aligned}$$

By means of the equation 4.5 we have that $\mathbb{P}(\ Q_n \mid X^{n-1}, Q^{n-1}\ ) = \mathbb{P}(\ Q_n \mid Q^{n-1}\ )$, from which it follows that

$$\begin{aligned} L(\ X \mid Q\ ) &= \mathbb{P}(\ X_n \mid \mathbb{X}_{n-1}, Q_n\ )\mathbb{P}(\ X^{n-1} \mid Q^{n-1}\ ) \\ &= \prod_{t=1}^{n} \mathbb{P}(\ X_t \mid \mathbb{X}_t, Q_t\ ) \end{aligned}$$

and the conditional $\log$-likelihood can be defined as it follows

Definition 4.3.1 *For our purpose we will define the conditional* $\log$*-likelihood as*

$$l(\ X \mid Q\ ) = \sum_{t=1}^{n} \log \mathbb{P}(\ X_t \mid \mathbb{X}_{t-1}, Q_t\ ). \tag{4.8}$$

Taking into account the hidden process we rewrite it as

$$l(\ X \mid Q\ ) = \sum_{t=1}^{n}\sum_{k=1}^{K} S_{t,k} \log \mathbb{P}(\ X_t \mid \mathbb{X}_t, Q_t = k\ ). \tag{4.9}$$

If we were to assume the Z_t are i.i.d. standard normal random variables we would have

$$X_t \quad \text{is} \quad \mathcal{N}(\ m_k(\ \mathbb{X}_{t-1}\ ), \sigma_k^2(\ \mathbb{X}_{t-1}\ )\ )$$

for $S_{t,k} = 1$. Therefore, under the normality assumption of the residuals one can rewrite the conditional log-likelihood as follows

$$l(\ X \mid Q\ ) = \sum_{t=1}^{n}\sum_{k=1}^{K} S_{t,k} \left(\log \frac{1}{\sqrt{2\pi\sigma_k^2(\ \mathbb{X}_{t-1}\ )}} - \frac{(\ X_t - m_k(\ \mathbb{X}_{t-1}\ )^2}{2\sigma_k^2(\ \mathbb{X}_{t-1}\ )} \right) \tag{4.10}$$

Now, assuming that $m_k(\ \mathbb{X}_{t-1}\ )$ and $\sigma_k(\ \mathbb{X}_{t-1}\ )$ can be approximated by suitable Feedforward Networks with finite number of hidden neurons, we want to find the asymptotic properties of the parameter estimates although the S_t are unknown. For this purpose we need some assumptions that will be clarified in the next section.

4.3.1 Consistency of the Parameter Estimates

In this section we shall assume that the residuals are i.i.d. normal and make use of the universal approximation property of the Neural Networks as presented in the previous chapter. We will also assume that Feedforward Networks with finite numbers of hidden neurons can suitably approximate the autoregressive and volatility functions. Under these considerations, we need to find the asymptotic properties of the parameter estimates of the pseudo conditional log-likelihood, i.e. with

$\theta = (\theta_1, \beta_1, \cdots, \theta_K, \beta_K)$

$$\hat{l}(\,\theta\,) \;=\; \sum_t \left(\log \frac{1}{\sqrt{2\pi \sum_k S_{t,k} f_{\sigma_k}(\,\mathbb{X}_{t-1} \mid \beta_k\,)}} - \frac{\sum_k S_{t,k}(\,X_t - f_{m_k}(\,\mathbb{X}_{t-1} \mid \theta_k\,)\,)^2}{2 \sum_k S_{t,k} f_{\sigma_k}(\,\mathbb{X}_{t-1} \mid \beta_k\,)} \right), \qquad (4.11)$$

where (in the conditional log-likelihood) we have replaced the m_k and σ_k by their network estimates f_{m_k} and f_{σ_k} respectively.
Under these hypotheses, we follow a realistic approach. This means we will assume the possibility for the model to be misspecified.

Comments on Misspecification

In this context we allow for misspecification. Indeed, there are various sources of misspecification, e.g. the system can not be uniquely determined or not even be determined at all. Still, if we assume that the system to be uniquely determined, the normality assumption made on the residuals can be violated. If one considers that the normality assumption is violated, one can refer to Amemiya ([1], Section 8.2.3) to notice that under this consideration the consistency proof can be derived just in some very specific situations, that means, the parameter estimates are in general not consistent. Following this observation, for our problem, we will consider the normality assumption is not violated, but still we will allow the system to be misspecified. By doing so we will make use of the result by Pötscher and Prucha [76] chapter 14 to prove the consistency of the parameter estimate assuming that the volatility and autoregressive functions can be approximated by suitable Feedforward Networks.

Stability and Identifiability

The concept of stochastic stability introduced by Bierens [4], the ν-stability in L^2 used by Billingsley [8], the mixing conditions or the ergodic property usually considered for the proof of some asymptotic results can all be regarded as stability conditions. Pötscher and Prucha [76] have proposed a more general concept, the L_p-approximability that, unlike the previous concepts, induces stability conditions. In the coming lines we recall the L_0-approximability definition and one can refer to their book for more comprehensive details on the general concept of L_p-approximability.

Definition 4.3.2 *(L_0-approximability)*
Let $\{v_t, t \in \mathbb{N}\}$ and $\{e_t, t \in \mathbb{Z}\}$ be stochastic processes defined on $(\,\Omega, \mathcal{F}, \mathbb{P}\,)$ that take their values in $\mathbb{R}^{p_v}$ and $\mathbb{R}^{p_e}$, respectively. Then, the process $\{v_t, t \in \mathbb{N}\}$ is

called L_0-approximable by the basis process $\{e_t, t \in \mathbb{Z}\}$ if there exist measurable functions $h_t^m : \mathbb{R}^{(\,2m+1\,)p_e} \longrightarrow \mathbb{R}^{p_v}$ such that for every $\delta > 0$ we have

$$\limsup_{n\longrightarrow\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{P}(\ |v_t - h_t^m(\ e_{t+m}, \cdots, e_{t-m}\ )| > \delta\ ) \longrightarrow 0 \text{ as } m \longrightarrow \infty$$

Given this definition, we need to set some assumptions that will help us to work in the framework of Pötscher and Prucha [76].

A. 4.2 *(Mixing Condition)*
The process $\{(\ X_t, S_t\ )\}$ is α-mixing.

This assumption will essentially help us to work in the L_p-approximability framework, in the sense that such an assumption implies that $\{(\ X_t, S_t\ )\}$ is trivially L_0-approximable by itself. Beside this assumption we would also like to make sure that the parameters of the model can be uniquely determined. Therefore, we need some identifiability conditions for the Network functions and the parameters and the Markov Chain as well. Thus, we have to make sure (that during the estimation) the parameters are not going to stick around a flat region.

A. 4.3 *(Identifiability Conditions)*

1. *The activation functions of the networks are antisymmetric, i.e. $\Psi(\ -u\ ) = -\Psi(\ u\ )$. Moreover, for*

$$f_H(\ x\ ) = \nu_o + \sum_{h=1}^{H} \nu_h \Psi\left(\omega_{0,h} + \sum_{i=1}^{p} \omega_{i,h} X_i\right),$$

we also assume

$$\nu_1 \geq \nu_2 \geq \cdots \geq \nu_H \geq 0 \tag{4.12}$$

2. *The initial distribution is ordered as*

$$\pi_1 \geq \pi_2 \geq \cdots \geq \pi_K > 0. \tag{4.13}$$

Remark 4.4 *Assumption **A.** 4.3 guarantees the identifiability and uniqueness of the parameters. Indeed **A.** 4.3, 1 guarantees the identifiability of the network parameter except on a set of measure 0 in the parameter space, i.e. except the case where $\nu_i = \nu_j$ for some i, j or $\nu_H = 0$.*
*If assumption **A.** 4.3, 2 is not required, by just renumbering the states one would produce different parameters but for the same process and the identifiability uniqueness*

will be violated.
Unlike what was done up to this chapter, as far as the parameters are concerned we have at least checked the consistency of their estimates. Having stated the above assumptions we can now move in that direction.

Moment Assumptions

A. 4.5

$$\mathbb{E}|X_t|^\gamma < \infty \quad \textit{for some} \quad \gamma > 0 \tag{4.14}$$
$$\mathbb{E}\|S_t\|^\gamma < \infty \quad \textit{for some} \quad \gamma > 0 \tag{4.15}$$

The moment assumption on S_t is obviously satisfied since independently of the choice of the norm we have $\|S_t\| = 1$.

Regularity Assumptions

Here we present some assumption on the parameter set, the activation function and the network functions used for the approximation of the volatility functions.

A. 4.6 *(Regularity Assumptions)*

1. *Assume that $A \subset \mathbb{R}^{2K(H(p+2)+1)}$, the set of all parameters is compact.*
2. *Assume that the activation functions (of the autoregressive and volatility functions) are continuously differentiable on $\mathbb{R}$ and bounded by 1 in absolute value.*
3. *$\exists \quad \varepsilon > 0$ such that $f_{\sigma_k}(u \mid \beta_k) \geq \varepsilon \quad \forall \quad u \in \mathbb{R}^p \quad \beta_k$ component of $\theta \in A$.*

Asymptotic Properties of the Parameter Estimates

Considering all the above assumptions, we are now in possession of all the ingredients to state and prove the consistency result for the parameter estimates under this setting. What we summarize in the following theorem.

Theorem 4.7 *Let us define*

$$\begin{aligned} R_n(\theta) &= -\frac{1}{n}\hat{l}(\theta) \\ &= \frac{1}{n}\sum_{t=1}^{n} q_t(z_t, \theta) \end{aligned}$$

be our objective function with $\theta = (\theta_1, \beta_1, \cdots, \theta_K, \beta_K)$ *where* $z_t = (X_t, \mathbb{X}_{t-1}, S_t)$ *Let also define*

$$\begin{aligned} \bar{R}_n(\ \theta\ ) &= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}q_t(\ z_t, \theta\ ) \\ &= \mathbb{E}q_t(\ z_t, \theta\ ). \end{aligned}$$

Assuming that assumptions **A.***4.2 to* **A.***4.6 hold, it follows that*

$$\sup_A |R_n(\ \theta\ ) - \bar{R}_n(\ \theta\ )| \underset{i.p.}{\longrightarrow} 0 \text{ as } n \longrightarrow \infty$$

and $\{\bar{R}_n : n \in \mathbb{N}\}$ *is equicontinuous on A. Additionally, assume* $\bar{\theta}_0$ *is the unique minimizer of* $\bar{R}_n$ *i.e.*

$$\bar{\theta}_0 = \inf_A \bar{R}_n,$$

and let $\hat{\theta}_n$ *be any sequence of minimizer of* R_n*, i.e. satisfying,*

$$\hat{\theta}_n = \inf_A R_n(\ \theta\ ).$$

It follows that $\hat{\theta}_n$ *is consistent for* $\bar{\theta}_0$*, i.e.* $|\hat{\theta}_n - \bar{\theta}_0| \underset{i.p.}{\longrightarrow} 0$ *as* $n \longrightarrow \infty$.

Proof: Following the approach of Pötscher and Prucha [76], Chapter 14.1, we pretend that $\{X_t\}$ is generated by

$$F_t(X_t, \mathbb{X}_{t-1}, S_t, \theta) = Z_t \tag{4.16}$$

where Z_t are i.i.d. standard normal random variables and

$$F_t(x, u, s, \theta) = \sum_k s_k \frac{x - f_{m_k}(u|\theta_k)}{f^{1/2}_{\sigma_k}(u|\beta_k)}. \tag{4.17}$$

We stress that $\{X_t\}$ is still a general mixture process of the form 4.1, but in estimating $\theta = (\theta_1, \beta_1, \cdots, \theta_K, \beta_K)$ we pretend that the residuals are normal and the autoregressive and volatility functions are neural network of the form $f_{m_k}(u|\theta_k)$ and $f^{1/2}_{\sigma_k}(u|\beta_k)$ resp. In general, therefore, we use a misspecified model for estimation. We remark that

$$\frac{\partial}{\partial x} F_t(x, u, s, \theta) = \sum_k s_k \frac{1}{f^{1/2}_{\sigma_k}(u|\beta_k)}.$$

Now, using this notation and Equation 4.11, we have

$$q_t(z_t, \theta) = \frac{1}{2}\log(2\pi) - \log|\frac{\partial}{\partial x} F_t(X_t, \mathbb{X}_{t-1}, S_t, \theta)| + \frac{1}{2} F_t^2(X_t, \mathbb{X}_{t-1}, S_t, \theta)$$

i.e. it is the form used in Pötscher and Prucha [76], Chapter 14.1, with $\Sigma = var(Z_t) = 1$ given. Therefore, we can apply Theorem 14.1 of Pötscher and Prucha [76], and have to check if the conditions of this consistency results are satisfied.
Obviously, because of the choice of activation functions, $F_t(\ x, u, s, \theta\ )$ and $\frac{\partial}{\partial x}F_t(\ x, u, s, \theta\ )$ are continuous functions that depend on t only via u_t. Hence $\{f_t : t \in \mathbb{N}\}$ and $\left\{\log(\ \frac{\partial}{\partial x}F_t\ ) : t \in \mathbb{N}\right\}$ are equicontinuous on $U \times A$ where $U = \mathbb{R}^{p+1} \times \mathcal{K}$ and $A \subseteq \mathbb{R}^l$ compact represents the set of parameters of our model.
By assumption we have

$$\begin{aligned}\mathbb{E}|X_t|^\gamma < \infty \quad &\text{for some} \quad \gamma > 0\\ \mathbb{E}\|S_t\|^\gamma < \infty \quad &\text{for some} \quad \gamma > 0\end{aligned}$$

and since $F_t(\ x, u, s, \theta\ )$ is independent of t, it follows that

$$\sup_t |F_t(\ x, u, s, \theta\ )| < \infty \quad \text{for all} \quad (\ x, u, s, \theta\ ) \in U \times A \tag{4.18}$$

Now consider $\varepsilon > 0$ give by **A.** 4.6 such that

$$f_{\sigma_k}(\ \mathbb{X}_{t-1} \mid \beta_k\ ) \ \geq \ \varepsilon \quad \forall \quad \mathbb{X}_{t-1} \quad \beta_k,$$

and it follows that

$$|F_t(\ x, u, s, \theta\ )| \leq \frac{1}{\varepsilon}\sum_k |x - f_{m_k}(\ u \mid \theta_k\ )| \quad \text{and}$$

$$\log(\ \varepsilon\ ) \leq \log(\ \frac{\partial}{\partial x}F_t\ ) \leq \sum_k \log(\ f_{\sigma_k}(\ u \mid \beta_k\ )\ )$$

and because A is compact it follows that

$$\exists M > 0, \text{ such that, } \forall u,\ \theta \in A,\ \forall\ k$$

$$\begin{aligned}&-M \leq f_{m_k}(\ u \mid \theta_k\ ) \leq M \text{ i.e.}\\ &x - M \leq x - f_{m_k}(\ u \mid \theta_k\ ) \leq x - M.\end{aligned}$$

Hence

$$|X_t - f_{m_k}(\ \mathbb{X}_{t-1} \mid \theta_k\ )| \leq |X_t - M| + |X_t - M|$$

and therefore,

$$\sum_k |X_t - f_{m_k}(\ \mathbb{X}_{t-1} \mid \theta_k\ )| \leq K(\ |X_t - M| + |X_t - M|\ ).$$

Finally

$$\mathbb{E}\sup_A(\ \sum_k |X_t - f_{m_k}(\ \mathbb{X}_{t-1} \mid \theta_k\ )|\ )^{2\gamma+2} \leq const(\ \mathbb{E}|X_t - M|^{2\gamma+2} + \mathbb{E}|X_t - M|^{2\gamma+2}\ ).$$

Analogously, $\exists\ M_\sigma > 0$ such that for all $u,\ \theta \in A$

$$\mathbb{E} \sup_A |\log(\ \frac{\partial}{\partial x} F_t\ )| \leq const(\ |\log \varepsilon| + |\log M_\sigma|\ ).$$

By just assuming that $\mathbb{E}|X_t|^{2\gamma+2} < \infty$ we can conclude that

$$\sup_n \frac{1}{n} \sum_{t=1}^{n} \mathbb{E} \sup_A (\ \sum_k |X_t - f_{m_k}(\ \mathbb{X}_{t-1} \mid \theta_k\ )|\ )^{2\gamma+2} \ < \ \infty \tag{4.19}$$

and

$$\sup_n \frac{1}{n} \sum_{t=1}^{n} \mathbb{E} \sup_A |\log(\ \frac{\partial}{\partial x} F_t\ )|^{1+\gamma} \ < \ \infty. \tag{4.20}$$

Given all the previous calculus and considerations we shall conclude the proof by considering the compactness of A, the equicontinuity of $F_t(\ x, u, s, \theta\ )$ and $\frac{\partial}{\partial x} F_t(\ x, u, s, \theta\ )$ on $U \times A$ and equations 4.14 to 4.20 to make sure that assumption 14.1 in chapter 14 of the book by Pötscher and Prucha [76] holds. Assumption 14.2 from the same book holds by the mixing assumption on $\{(\ X_t, S_t\ )\}$. Moreover, assuming the identifiability uniqueness of $\bar{\theta}_0$, which is guaranteed by the assumption on the identifiability, the proof of this theorem follows as special case of Theorem 14.1 in [76]. ∎

4.4 EM Algorithm

Baum et al ([3], 1970) proposed an elegant procedure to compute $\mathbb{P}(\ X^n \mid \theta\ )$ and Dempster, Lair and Rubin (1977) introduced the so-called **Expectation Maximization** algorithm to maximize this probability. This last proposal can be regarded as an extension of the Forward-Backward procedure.

4.4.1 Generalities on EM Algorithms

The EM algorithm is very popular for, e.g. one can take into consideration the simplicity and the generality of the underlying theory. Moreover this procedure can be applied in various contexts. In this section we focus on its most classical description, that we consider it definition for missing or hidden data models.
For this purpose we consider the observed data denoted by X and the hidden data S. We then defined the conditional probability distribution of the extended data model given the vector parameter θ, $f_e(X, S \mid \theta)$ from which we derive the marginal probability distribution of the observed data model

$$f_d(X \mid \theta) = \int f_e(X, S \mid \theta) dS.$$

The goal of the EM algorithm is then to maximize the observed data $\log$-likelihood function, i.e.

$$L_d(\theta) = \log\left(f_d(X \mid \theta)\right).$$

This problem is not addressed directly, with the EM algorithm, one will rather solve iteratively the $\log$-likelihood of the extended data model

$$L_e(\theta) = \log\left(f_e(X, S \mid \theta)\right),$$

which is a random variable due to the hidden observations S. More precisely, let $\hat{\theta}_m$ denote the value of the estimator of θ on the iteration m of the EM algorithm we then compute in the E-step of this iteration

$$\mathbf{Q}(\theta, \hat{\theta}_m) = \mathbb{E}\left(L_e(\theta)\right))$$

where the expectation is computed with respect to $\hat{\theta}_m$. In the M-step of the same iteration, we compute

$$\hat{\theta}_{m+1} = \arg\max_{\theta} \mathbf{Q}(\theta, \hat{\theta}_m).$$

The algorithm is started with an initial vector value $\hat{\theta}_0$ of the parameter θ and the E-step and M-step are iterated until some stopping criterion is satisfied. In general, with the EM algorithm $L_e(\theta)$ is non decreasing, i.e.

$$L_e(\hat{\theta}_{m+1}) \geq L_e(\hat{\theta}_m), \quad \text{for} \quad m = 0, 1, \cdots.$$

For more details on this issue and some other classical properties of the EM algorithm, one can refer to Baum et al [3] or have a look to [31].

4.4.2 Forward-Backward Procedure

Let us first announce that, in general, all over this section the probabilities are defined conditioned on the parameter vector θ, even in the cases this is not explicitly specified. Indeed, it will happen that we avoid this specification just to simplify the writing. We now return to our mixture models.

Forward Procedure

Let α_i^t be the joint probability of having the observation from time $-p+1$ to t and being in state i at time t.

$$\begin{aligned} \alpha_i^t &= \mathbb{P}(X_{-p+1}, \cdots, X_1, \cdots, X_t, Q_t = i \mid \theta) \\ &= \mathbb{P}(X_{-p+1}, \cdots, X_1, \cdots, X_t \mid Q_t = i, \theta)\mathbb{P}(Q_t = i),\ 1 \leq t \leq n; \end{aligned} \tag{4.21}$$

where $\mathbb{P}(x \mid Q_t = i, \theta)$ is the conditional density of $(X_{-p+1}, \cdots, X_1, \cdots, X_t)'$ given $Q_t = i$ and θ denotes the model parameters. It follows that the density the

complete sequence of observations is given by the sum over all states at the end (n) of the sequence, i.e.

$$\mathbb{P}(\ X^n\ \mid\ \theta\ ) = \sum_{i=1}^{K} \alpha_i^n \tag{4.22}$$

The surprising about this representation is its computational complexity. Rather than being exponential in n, it is only linear in time since α_i^n can be computed recursively.

$$\begin{aligned}
\alpha_j^{t+1} &= \mathbb{P}(X_{-p+1}, \cdots, X_1, \cdots, X_t, X_{t+1}, Q_{t+1} = j\ \mid \theta\ ) \\
&= \mathbb{P}(\ X_{t+1}\ \mid X^t, Q_{t+1} = j\ )\mathbb{P}(\ X^t, Q_{t+1} = j\ ) \\
&= \mathbb{P}(\ X_{t+1}\ \mid X^t, Q_{t+1} = j\ ) \sum_{i=1}^{K} \mathbb{P}(\ X^t, Q_{t+1} = j, Q_t = i\ ) \\
&= \mathbb{P}(\ X_{t+1}\ \mid X^t, Q_{t+1} = j\ ) \sum_{i=1}^{K} \mathbb{P}(\ Q_{t+1} = j\ \mid X^t, Q_t = i\ )\mathbb{P}(\ X^t, Q_t = i\ ) \\
&= \mathbb{P}(\ X_{t+1}\ \mid \mathbb{X}_{t+1}, Q_{t+1} = j\ ) \sum_{i}^{K} \mathbb{P}(\ Q_{t+1} = j\ \mid Q_t = i\ )\mathbb{P}(\ X^t, Q_t = i\ ) \\
&= b_j^{t+1} \left[\sum_{i=1}^{K} a_{ij} \alpha_i^t \right].
\end{aligned} \tag{4.23}$$

This sequence can be initialized with

$$\alpha_i^1 = \pi_i b_i^1. \tag{4.24}$$

This step is called the forward procedure, given the initial values of π_i and b_1^i .

Backward Procedure

In the same way as above we will define β_i^t (the Backward variable) as the conditional density of observing $X_s, s = t+1, \cdots, n$ given the state i at time t and the past realizations of the process $\mathbb{X}_{t+1}$

$$\begin{aligned}
&\beta_i^t \\
&= \mathbb{P}(\ X_{t+1}, \cdots, X_n\ \mid \mathbb{X}_{t+1}, Q_t = i\ ) \quad \forall\ i \\
&= \sum_j \mathbb{P}(\ X_{t+1}, \cdots, X_n, Q_{t+1} = j\ \mid \mathbb{X}_{t+1}, Q_t = i\ ) \\
&= \sum_j \mathbb{P}(\ X_{t+2}, \cdots, X_n, \mid \mathbb{X}_{t+2}, Q_{t+1} = j\ )\mathbb{P}(\ X_{t+1}\ \mid \mathbb{X}_{t+1}, Q_{t+1} = j\ )\mathbb{P}(\ Q_{t+1} = j\ \mid Q_t = i\ ) \\
&= \sum_j a_{ij}\, b_j^{t+1}\, \beta_j^{t+1},
\end{aligned} \tag{4.25}$$

for $t = n-1, n-2, \cdots, 1$ and the recursion starts with $\beta_i^n = 1$.
Obviously, we derive

$$\mathbb{P}(X^n, Q_t = i) = \alpha_i^t \beta_i^t \tag{4.26}$$

Auxiliary Variables

Since the state variables $S_{t,k}$ are unknown and random we would like to replace them by their conditional expectations. To this end we compute the posterior probability of being in state i at time t given the entire sequence of observations and the parameters of the model.

$$\begin{aligned}\gamma_i^t &= \mathbb{P}(\, Q_t = i \mid X^n \,) \\ &= \frac{\mathbb{P}(\, Q_t = i, X^n \,)}{\mathbb{P}(\, X^n \,)} \\ &= \frac{\mathbb{P}(\, Q_t = i, X^n \,)}{\sum_{k=1}^K \mathbb{P}(\, Q_t = k, X^n \,)} \\ &= \frac{\alpha_i^t \beta_i^t}{\sum_{k=1}^K \alpha_k^t \beta_k^t}. \end{aligned} \tag{4.27}$$

Finally, the joint conditional probability $\xi_{ij}^{t,t+1} = \mathbb{P}(\, Q_t = i, Q_{t+1} = j \mid X^n \,)$ of Q_t and Q_{t+1} is given as follows

$$\begin{aligned}\xi_{ij}^{t,t+1} &= \mathbb{P}(\, Q_t = i, Q_{t+1} = j \mid X^n \,) \\ &= \frac{\mathbb{P}(\, Q_t = i, Q_{t+1} = j, X^n \,)}{\mathbb{P}(\, X^n \,)} \\ &= \frac{\mathbb{P}(\, Q_t = i, Q_{t+1} = j, X^n \,)}{\sum_{k=1}^K \alpha_k^t \beta_k^t} \\ &= \frac{a_{i,j} \alpha_i^t b_j^{t+1} \beta_j^{t+1}}{\sum_{k=1}^K \alpha_k^t \beta_k^t}, \end{aligned} \tag{4.28}$$

since

$$\begin{aligned} & \mathbb{P}(\, Q_t = i, Q_{t+1} = j, X^n \,) \\ = \;& \mathbb{P}(\, X_{t+2}, \cdots, X_n \mid Q_t = i, Q_{t+1} = j, X^{t+1} \,) \mathbb{P}(\, Q_t = i, Q_{t+1} = j, X^{t+1} \,) \\ = \;& \mathbb{P}(\, X_{t+2}, \cdots, X_n \mid Q_{t+1} = j, \mathbb{X}_{t+2} \,) \mathbb{P}(\, Q_t = i, Q_{t+1} = j, X^{t+1} \,) \\ = \;& \beta_j^{t+1} \mathbb{P}(\, X_{t+1} \mid Q_t = i, Q_{t+1} = j, X^t \,) \mathbb{P}(\, Q_t = i, Q_{t+1} = j, X^t \,) \\ = \;& \beta_j^{t+1} \mathbb{P}(\, X_{t+1} \mid Q_{t+1} = j, \mathbb{X}_{t+1} \,) \mathbb{P}(\, Q_{t+1} = j \mid Q_t = i, X^t \,) \mathbb{P}(\, Q_t = i, X^t \,) \\ = \;& a_{i,j}\, \alpha_i^t\, b_j^{t+1}\, \beta_j^{t+1}. \end{aligned}$$

Having this auxiliary variable one can compute the estimates of the transition probabilities and the initial distribution of the chain i.e.

$$\begin{aligned}\hat{a}_{ij} &= \frac{\text{Expected number of transitions from state i to state j}}{\text{Expected number of transitions from i to anywhere}} \\ &= \frac{\sum_t \hat{\xi}_{ij}^{t,t+1}}{\sum_t \hat{\gamma}_i^t} \end{aligned}$$

and

$$\hat{\pi}_i = \frac{1}{n}\sum_t \hat{\gamma}_i^t$$

where $\hat{\gamma}_i^t, \hat{\xi}_{ij}^{t,t+1}$ are estimates of $\gamma_i^t, \xi_{ij}^{t,t+1}$ calculated during the expectation step of the EM-iteration. To compute the conditional expectation of the state variables $S_{t,k}$ as we did previously is somehow "cheating". Indeed, for their computation, we did not only use the past information but the entire training set. Therefore, their estimates are non causal which in a pure statistical framework is considered as a drawback of this procedure. But on the other hand we have to observe that we did not use future data in the strict sense that the training set is always at our disposal. However, a causal version of these conditional could have been obtained through a few computation stage, what we summarize as it follows,

$$\begin{aligned}
& \mathbb{P}(\, Q_t = k \mid X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,) \\
= & \frac{\mathbb{P}(\, Q_t = k, X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,)}{\mathbb{P}(\, X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,)} \\
= & \frac{\mathbb{P}(\, Q_t = k, X_t, \cdots, X_1, \cdots, X_{-p+1}\,)}{\sum_{j=1}^K \mathbb{P}(\, Q_t = j, X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,)} \\
= & \frac{\sum_{i=1}^K \mathbb{P}(\, Q_t = k, Q_{t-1} = i, X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,)}{\sum_{j=1}^K \sum_{i=1}^K \mathbb{P}(\, Q_t = j, Q_{t-1} = i, X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,)} \\
= & \frac{\sum_{i=1}^K \alpha_i^{t-1} a_{i,k}}{\sum_{j=1}^K \sum_{i=1}^K \alpha_i^{t-1} a_{i,j}},
\end{aligned}$$

from which we derive

$$\begin{aligned}
\mathbb{E}(\, S_{t,k} \mid X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,) &= \mathbb{P}(\, S_{t,k} = 1 \mid X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,) \\
&= \mathbb{P}(\, Q_t = k \mid X_{t-1}, \cdots, X_1, \cdots, X_{-p+1}\,) \\
&= \frac{\sum_{i=1}^K \alpha_i^{t-1} a_{i,k}}{\sum_{j=1}^K \sum_{i=1}^K \alpha_i^{t-1} a_{i,j}}.
\end{aligned}$$

Remark that from the Forward-Backward Procedure we can derive the estimates of the state variables and additionally obtain those of the transition probability matrix and of the initial distribution as well. Therfore we can say that we have a first step optimization in which the transition probability matrix and the initial distribution are the byproducts. Now we need to complete the estimation procedure in order to obtain a full set of parameters for the model.

4.4.3 Maximization

In this section we can consider the state variables $S_{t,k}$ to be known, i.e., more precisely, we replace them with their conditional expectations $\hat{S}_{t,k}$ as estimated via the

γ_t^k (compare equation 4.27) and minimize the following equation w.r.t. the network parameters θ.

$$G(\ \theta\ ) \ = \ \sum_{t=1}^{n}\sum_{k=1}^{K}\hat{S}_{t,k}\left(\ \log\sqrt{f_{\sigma_k}(\ \mathbb{X}_{t-1},\beta_k\ )}+\frac{(\ X_t-f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ )^2}{2f^2_{\sigma_k}(\ \mathbb{X}_{t-1},\beta_k\ )}\right),$$

where in the last formulation the m_k, σ_k are replaced by suitable Feedforward Networks. Their first order derivatives w.r.t. θ can be written as it follows

$$\frac{\partial G(\ \theta)}{\partial\theta_{k,i}} \ = \ -\sum_{t=1}^{n}\hat{S}_{t,k}\frac{\partial f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ )}{\partial\theta_{k,i}}\frac{(\ X_t-f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ )}{f_{\sigma_k}(\ \mathbb{X}_{t-1},\beta_k\ )}$$

and

$$\frac{\partial G(\ \theta\ )}{\partial\beta_{k,j}} \ = \ \frac{1}{2}\sum_{t=1}^{n}\hat{S}_{t,k}\frac{\partial f_{\sigma_k}(\ \mathbb{X}_{t-1},\beta_k\ )}{\partial\beta_{k,j}}\frac{1}{f_{\sigma_k}(\ \mathbb{X}_{t-1},\beta_k\ )}\left(\ 1-\frac{(\ X_t-f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ )\ )^2}{f_{\sigma_k}(\ \mathbb{X}_{t-1},\beta_k\ )}\right).$$

Numerically, we can retrieve the network's parameters by using a stochastic approximation algorithm as e.g. the stochastic gradient algorithm.
Now, let us focus on a special case where we have to consider the volatility functions to be constant but different, i.e. $\sigma_k^2(\ x\ ) = \sigma_k^2$. In that case, we do not need a neural network f_{σ_k}, but estimate the parameter σ_k^2 directly. It follows that by solving

$$\frac{\partial G(\ \theta\ )}{\partial\sigma_k^2} \ = \ 0$$

we derive

$$\sigma_k^2 \ = \ \frac{\sum_{t=1}^{n}\hat{S}_{t,k}(\ X_t-f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ )\ )^2}{\sum_{t=1}^{n}\hat{S}_{t,k}}.$$

Intuitively, that is just the usual residual variance estimate of the k/th subsample in the mixture models.
Similarly solving

$$\frac{\partial G(\ \theta\ )}{\partial\theta_{k,i}} \ = \ 0$$

is equivalent to solving

$$\sum_{t=1}^{n}\hat{S}_{t,k}\frac{\partial f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ )}{\partial\theta_{k,i}}(\ X_t-f_{m_k}(\ \mathbb{X}_{t-1},\theta_k\ ) \ = \ 0$$

For this special case we need to observe that for the σ_k^2 we have obtained an analytical formula; but this representation depends on the unknown autoregressive functions which under our considerations are parametric functions. Once more we can retrieve these parameters by using a stochastic gradient algorithm.

4.4.4 An Adaptation of the Expectation Maximization Algorithm

The procedures we presented in the Forward-Backward Procedure and the maximization steps can be summarized in a version of the well-known EM-algorithm, which we will call EM-Algorithm for GMAR-ARCH models.

Algorithm 4.8 *(EM-Algorithm for GMAR-ARCH models)*

1. *Set* $m = 0$ *and choose the initial value for the parameters* $\hat{\theta}_0$
2. *(Expectation or* ***E****-Step)*
   Assume that the parameters of the model are known, i.e, set $\theta = \hat{\theta}$ *and compute (for each time instant* t *) the forward variables* α_k^t *and the backward variables* β_k^t*; consequently the auxiliary variables* γ_i^t *and* $\xi_{ij}^{t,t+1}$.
3. *(Maximization or* ***M****-Step)*
   Consider the variables obtained in the ***E****-Step and maximize* $-G(\ \theta\ )$
4. *Replace* m *by* $m + 1$ *and repeat the procedure starting from the* ***E****-Step until a stopping criterion is satisfied.*

At this point we are able to talk in terms of consistency of the parameter estimates and their numerical estimation procedure as well. Nevertheless the less they are still many questions that we can address, e.g. how can we determine the optimal (most likely) hidden state sequence for our model given a sequence of observed outputs? To answer this question we would for example like to adapt the well-known Viterbi algorithm to our context.

4.5 Viterbi Algorithm

The Viterbi is an algorithm to compute the optimal (most likely) state sequence in a Hidden Markov Model given a sequence of observed outputs. It is based on the maximization of the single best state sequence and it is based on the dynamic programming method. In this case to find the single best state sequence $\{S_1, S_2, \cdots, S_n\}$ for the observations $\{X_1, X_2, \cdots, X_n\}$ we define

$$\delta_t(\ i\ ) = \max_{S_1,\cdots,S_{t-1}} \log \mathbb{P}(\ S_1, S_2, \cdots, S_{t,i} = 1, X_{1-p}, X_{2-p}, \cdots, X_t),$$

i.e. $\delta_t(\ i\ )$ is the highest $\log$-probability along a single path, at time t, which accounts for the first t observations and ends in state i. By induction we have

$$\begin{aligned}
\delta_{t+1}(\ j\ ) &= \max_{S_1,\cdots,S_t} \log \mathbb{P}(\ S_1, S_2, \cdots, S_{t+1,j} = 1, X_{1-p}, \cdots, X_2, \cdots, X_{t+1}\ ) \\
&= \max_{S_1,\cdots,S_t} \log \{\mathbb{P}(\ S_{t+1,j} = 1 \mid S_{t,i} = 1\ )\mathbb{P}(\ X_{t+1} \mid S_{t+1,j} = 1, \mathbb{X}_t\ ) \\
&\qquad\qquad \mathbb{P}(\ S_1, \cdots, S_{t,j} = 1, X_{1-p}, \cdots, X_t\ )\}\,,
\end{aligned}$$

i.e.

$$\delta_{t+1}(\ j\ ) = \max_i(\ \delta_t(\ i\ ) + \log a_{i,j}\ ) + \log b_j^{t+1}.$$

To retrieve the state sequence, we need to follow the trajectory delivered by the argument that maximized the previous equation for each t and j. We will achieve it via an auxiliary variable $\psi_t(\ j\ )$ and the complete procedure is written as follows

1. Initialization:

$$\begin{aligned}\delta_1(\ j\ ) &= \log \pi_j b_j^1 \quad 1 \leq j \leq K\\ \psi_1(\ j\ ) &= 0,\end{aligned}$$

2. Recursion:

$$\begin{aligned}\delta_t(\ j\ ) &= \max_{1\leq i\leq K}(\ \delta_{t-1}(\ i\ ) + \log a_{i,j}\ ) + \log b_j^t, \quad 2 \leq t \leq n, \quad 1 \leq j \leq K\\ \psi_t(\ j\ ) &= \arg\max_{1\leq i\leq K}(\ \delta_{t-1}(\ i\ ) + \log a_{i,j}\ ), \quad 2 \leq t \leq n, \quad 1 \leq j \leq K\end{aligned}$$

3. Termination:

$$\begin{aligned}\log P^* &= \max_{1\leq i\leq K}(\ \delta_n(\ i\ )\ )\\ q_n^* &= \arg\max_{1\leq i\leq K}(\ \delta_n(\ i\ )\ )\end{aligned}$$

4. Path (State Sequence) Backtracking:

$$q_t^* = \psi_{t+1}(\ q_{t+1}^*\ ), \quad t = n-1, n-2, \cdots, 1.$$

In this section, several aspects of the GMAR-ARCH models have been addressed. In particular we have defined a pseudo conditional log-likelihood for which we have established the consistency of the parameter estimates under some regularity assumptions. Additionally, we have also proposed a version of the EM algorithm that account for the numerical estimation of the model parameters. Last but not the least, we have proposed a modified version of the well-known Viterbi algorithm that allows us to compute the most likely state of the hidden process at each time instant.

5 Nonlinear Univariate Weighted Least Squares for Changepoint Analysis in Time Series Models

In this Chapter we are concerned with the estimation of the m_k from the model 2.2 at each time instant. For this purpose, we will proceed in two steps. We first suppose that there is no change in the dynamic of the observed time series, that is we first consider the case $K = 1$ and later we allow several dynamics within the observed time series. Under the consideration of several dynamics within the process, we also strengthen our assumption on the hidden process by considering the S_t as i.i.d. random variables. In this case we achieve our goal by using an appropriate weighted least squares that we need to solve. For the first case we simply deal with a classical problem of Nonlinear Least squares for which we have to find the asymptotic properties of the parameter estimates.

5.1 Nonlinear Least Squares

In this section, let us consider a process $\{X_t\}$ generated by a model of the type 2.2 with $K = 1$ and p a given positive integer, i.e.

$$X_{t+1} = m(\mathbb{X}_t) + \sigma(\mathbb{X}_t)\epsilon_{t+1} \tag{5.1}$$

where the ϵ_t are i.i.d. random variables with mean zero and constant variance. For sake of simplicity we will assume the latter to be equal to 1. Alternatively we can rewrite

$$X_{t+1} = m(\mathbb{X}_t) + Z_{t+1}, \tag{5.2}$$

with $Z_{t+1} = \sigma(\mathbb{X}_t)\epsilon_{t+1}$. Under these considerations, the conditional expectation of X_{t+1} given the past information is derived as follows,

$$\begin{aligned} m(x) &= \mathbb{E}(X_{t+1} \mid \mathbb{X}_t = x) \\ &= \mathbb{E}(X_{t+1} \mid (X_t, X_{t-1}, \cdots, X_{t-p+1}) = x). \end{aligned} \tag{5.3}$$

Similarly the conditional variance can be computed by

$$\mathbb{E}((X_t - m(\mathbb{X}_{t-1}))^2 \mid \mathbb{X}_t = x) = \sigma^2(x).$$

In this context, making use of the universal approximation property of the feedforward networks, as stated for the first time by White (compare e.g. [91]) or as we presented in the previous Chapter 3 under the main assumption of the change within the dynamic of the observed time series. We want to approximate m with a single hidden layer feedforward network containing H hidden neurons, $H \geq 1$. The output function can be written as

$$f(x_1, \cdots, x_p; \theta) = \nu_0 + \sum_{h=1}^{H} \nu_h \psi(\omega_h^{(0)} + \sum_{i=1}^{p} \omega_h^{(i)} x_i),$$

where $\theta = (\omega_h^{(0)}, \cdots, \omega_H^{(p)}, \nu_0, \cdots, \nu_H)'$ is the vector of weights, $\theta \in \mathbb{R}^l$, with $l = H(p+2)+1$, H is the number of hidden units and p the order of the observed nonlinear autoregressive ($NLAR$) process.
In general this output is different from the desired target, i.e. we allow the model to be misspecified[1]. Therefore, the need to find θ_0 for which $f(x, \theta)$ approximates X_{t+1} at best, in the sense that θ_0 is a the global minimizer of the expectation of the cost function, i.e.

$$\begin{aligned} \theta_0 &= \arg\min_{\theta \in \Theta_H} Q(\theta) \\ &= \arg\min_{\theta \in \Theta_H} \mathbb{E}(X_{t+1} - f(X_t, \cdots, X_{t-p+1}, \theta))^2, \end{aligned}$$

what we can rewrite as

$$\begin{aligned} Q(\theta) &= \mathbb{E}(m(X_t, \cdots, X_{t-p+1}) - f(X_t, \cdots, X_{t-p+1}, \theta))^2 \\ &\quad + \sigma^2(X_t, \cdots, X_{t-p+1}) \end{aligned}$$

However, the probability distribution here is usually unknown, therefore, we will rely on a Strong Law of Large Numbers or a version of the well-known Ergodic Theorem to derive some asymptotic properties. We shall also train the network to get the nonlinear least squares estimate θ_n of the weight vector. For this purpose, we consider a training set of random variables $X_t,\ t = -p+1, \cdots, n$ where $p \in \mathbb{N}$ is the order of a nonlinear AR (NLAR), and $n \in \mathbb{N}$ the sample training size. In fact, we need to find

$$\theta_n = \arg\min_{\theta \in \Theta} \frac{1}{n} Q_n(\theta), \tag{5.4}$$

where

$$\begin{aligned} Q_n(\theta) &= \sum_{t=1}^{n} \frac{(X_t - f(\mathbb{X}_{t-1}, \theta))^2}{2} \\ &= \sum_{t=1}^{n} q_t(\theta). \end{aligned}$$

For $\theta_0 \in \Theta$, one can prove that the parameter estimate is consistent and $\sqrt{n}(\theta_n - \theta_0)$ is asymptotically normal, with mean vector 0 and covariance matrix $(E(\nabla f(\mathbb{X}_{t-1}L, \theta_0) \nabla f(\mathbb{X}_{t-1}, \theta_0)'))^{-1}$. This is the case in the literature for i.i.d random variables and Yao [97] in the correctly specified case of the NLAR. Moreover, by an extension of the results by Klimko and Nelson [55] we derive these results even in the case where the model is misspecified.
Basically, to obtain the nonlinear least squares estimate, we need to solve " Least squares Equations" of the form

$$\frac{\partial Q_n}{\partial \theta_i}(\theta) = 0 \ \forall\ i = 1, \cdots, l.$$

[1] One can refer to the previous chapter for some comments on misspecification

For this purpose, we need to control the behavior of the second order term in the Taylor expansion of $Q_n(\theta)$ around θ_0 (given some neighborhood N_{θ_0}). In the remainder, all other neighborhoods will be included in N_{θ_0}. For $\delta > 0$, $\|\theta - \theta_0\| < \delta$, for some $\theta^*, 0 < \|\theta_0 - \theta^*\| < \delta, \theta^* = \theta_0 + \lambda(\theta - \theta_0)$ for some $\lambda \in (0, 1)$.

$$\begin{aligned} Q_n(\theta) &= Q_n(\theta_0) + (\theta - \theta_0)'\nabla Q_n(\theta_0) + \frac{1}{2}(\theta - \theta_0)'V_n(\theta - \theta_0) \\ &+ \frac{1}{2}(\theta - \theta_0)'T_n(\theta^*)(\theta - \theta_0) \end{aligned}$$

where

$$V_n = \left(\nabla^2 Q_n(\theta_0)\right), \quad T_n(\theta^*) = \nabla^2 Q_n(\theta^*) - V_n.$$

Before we move toward the proof of the asymptotic properties of this type of estimate, let us first present some preliminary results.

5.1.1 Preliminaries

Recall that

$$f(\mathbb{X}_t, \theta) = \nu_0 + \sum_{h=1}^{H} \nu_h \psi(\omega_h^{(0)} + \sum_{i=1}^{p} \omega_h^{(i)} X_{t-i+1}), \tag{5.5}$$

from which we obtain,

$$\frac{\partial q_t}{\partial \theta_i} = -\frac{\partial f(\mathbb{X}_t, \theta)}{\partial \theta_i}(X_{t+1} - f(\mathbb{X}_t, \theta)) \tag{5.6}$$

and it follows that

$$\nabla q_t = -\nabla f(\mathbb{X}_t, \theta)(X_{t+1} - f(\mathbb{X}_t, \theta)). \tag{5.7}$$

For the sake of simplicity we will present the 2^{nd} and 3^{rd} order partial derivative just in the case were moment assumptions of higher order are needed. In these extreme cases we should have for example

$$\frac{\partial f}{\partial \theta_i}(x_{t-1}, \cdots, x_{t-p}, \theta) = \begin{cases} 1 & \text{if } \theta_i = \nu_0 \\ \psi(\omega_h^{(0)} + \sum_{i=1}^{p} \omega_h^{(i)} x_{t-i+1}) & \text{if } \theta_i = \nu_h \\ \nu_h \psi'(\omega_h^{(0)} + \sum_{i=1}^{p} \omega_h^{(i)} x_{t-i+1}) & \text{if } \theta_i = \omega_h^{(0)} \\ x_{t-i+1}\nu_h \psi'(\omega_h^{(0)} + \sum_{i=1}^{p} \omega_h^{(i)} x_{t-i+1}) & \text{if } \theta_i = \omega_h^{(i)} \end{cases}$$

$$\frac{\partial^2 f}{\partial\theta_i\theta_j} = x_{t-i+1}x_{t-j+1}\nu_h\psi''(\omega_h^{(0)} + \sum_{i=1}^{p}\omega_h^{(i)}x_{t-i+1}) \text{ if } \theta_u = \omega_h^{(u)},$$

$$\frac{\partial^3 f}{\partial\theta_i\theta_j\theta_k} = x_{t-i+1}x_{t-j+1}x_{t-k+1}\nu_h\psi'''(\omega_h^{(0)} + \sum_{i=1}^{p}\omega_h^{(i)}x_{t-i+1}) \text{ if } \theta_u = \omega_h^{(u)}.$$

Given all the above mentioned details, we can present a set of weak assumptions that allows us to derive the consistency of the parameter estimates.

5.1.2 Consistency under Weak Assumptions

The main assumptions are now listed.

A. 5.1 *(Moment Assumptions)*

1. *Assume* $\{X_t\}$ *is (strictly) stationary,* α*-mixing and* $\mathbb{E}\,|X_t|^{2\gamma} < \infty$ *for some* $\gamma > 2$.
2. $\exists\, C_j > 0$ *such that* $\mathbb{E}(|Z_t|^j \mid \mathbb{X}_t = x) \leq C_j < \infty,\ \forall x$ *and for* $j = 1, \cdots, 4$.
3. $m(x)$ *is continuous and* $\exists\ \kappa \geq 0, a_i \geq 0$ *for* $i = 1, \cdots, p$ *such that* $|m(x)| \leq \sum_{i=1}^{p} a_i|x_{t-i+1}| + \kappa$.

A. 5.2 *(Regularity Assumption)*
The activation functions ψ *are* $\mathcal{C}^3$ *with bounded derivatives.*

A. 5.3 Q *has a unique global minimum at* θ_0 *which is an interior point of* Θ *(a compact subspace of* $\mathbb{R}^l$ *), and* $\nabla^2 Q(\theta_0) = A(\theta_0)$ *is positive definite.*

Observe that the regularity assumption implies that Q_n also satisfies some regularity assumption with respect to θ in some neighborhood of θ_0. Also, we need to remark that one can relax **A.**5.3 by assuming that Θ is an open subset of $\mathbb{R}^l$. Moreover, θ_0 do not need to be an interior point of the parameter set as it was pointed out in Amemiya [1].
Since we can derive the ergodic property from the mixing condition that we have assumed (compare the discussion in Hannan [44]), it will be redundant to assume a series to satisfy some mixing condition and to be ergodic. Hence, with the mixing assumption, the main theorem from the Ergodic Theory is granted.
Given these assumptions, let us now move toward our asymptotic results.

Proposition 5.4 *Let us consider* **A.***5.1 to* **A.***5.3 to hold, then we have the following:*

1. $$\lim_{n\to\infty} \frac{1}{n}\frac{\partial Q_n}{\partial\theta_i}(\theta_0) = 0 \text{ a.s., for } i = 1, \cdots, l,$$

2.

$$\lim_{n\to\infty}\frac{1}{n}V_n \to V \quad a.s.,$$

where V is a positive definite matrix.

3.

$$\lim_{n\to\infty}\sup_{\delta\to 0}\left(\frac{|T_n(\theta^*)_{i,j}|}{n\delta}\right) < \infty \text{ a.s. for all } \; i,j = 1,\cdots,l.$$

Remark 5.5 *First we observe that the above Proposition provides the conditions for the results by Klimko and Nelson [55] although they were established for i.i.d. random variables. Therefore, our results will be considered as an extension of these results in such a way that they are suitable for time series. To obtain this extension we need, for example, to assume that the observed process satisfies some mixing conditions. The mixing assumption can be considered as an asymptotic measure of independence. Also we have to assume higher moments to ensure the finiteness of the expectation up to that of the third partial derivatives of the cost function with respect to θ. Beside these considerations, the rest of the proof will follow just by the technique considered in [55]. Thus, we have to control the behavior of the Taylor expansion as announced earlier.*

Proof: The well-known Ergodic theorem for stationary time series can be considered as a corner stone of this proof.

1) Using **A.**5.1-**A.**5.3, the Ergodic Theorem for stationary time series and let consider i is given, we then obtain,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\frac{\partial Q_n}{\partial\theta_i}(\theta_0) &= \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{\partial q_t}{\partial\theta_i}\\ &= \mathbb{E}\frac{\partial q_t}{\partial\theta_i}(\theta_0)\\ &= \frac{\partial}{\partial\theta_i}Q(\theta_0) = 0 \; a.s.\end{aligned}$$

and the conclusion follows.

2) By a similar argument as above, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\nabla^2 Q_n(\theta_0) &= \nabla^2\mathbb{E}(X_{t+1} - f(\mathbb{X}_t,\theta_0))^2 \; a.s\\ &= \nabla^2 Q(\theta_0) = A(\theta_0),\end{aligned}$$

which is positive definite by **A.**5.3.

3) For the third part we have the following.

Let i, j be given, then we have

$$\begin{aligned} T_n(\ \theta^*\ )_{i,j} &= \frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta^*) - V_n(\ \theta_0)_{i,j} \\ &= \frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta^*) - \frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0) \\ &= (\ \theta^* - \theta_0)'\nabla\frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0) + \|\theta^* - \theta_0\| C(\ \theta^* - \theta_0). \end{aligned}$$

Where R represents the rest in the Taylor expansion of Q_n. Now, by an application of the triangle inequality and the Cauchy-Schwartz inequality, we derive

$$\begin{aligned} |T_n(\ \theta^*)_{i,j}| &= \left|(\ \theta^* - \theta_0\ )'\nabla\frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0) + \|\theta^* - \theta_0\| R(\ \theta^* - \theta_0)\right| \\ &\leq |(\ \theta^* - \theta_0)'\nabla\frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0)| + \|\theta^* - \theta_0\||R(\ \theta^* - \theta_0)| \\ &\leq \|\theta^* - \theta_0\|\ \|\nabla\frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0)\| + \|\theta^* - \theta_0\||R(\ \theta^* - \theta_0)|. \end{aligned}$$

It then follows that

$$\left|\frac{T_n(\ \theta^*)_{i,j}}{n\delta}\right| \leq \|\frac{1}{n}\nabla\frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0)\| + |R(\ \theta^* - \theta_0)|. \tag{5.8}$$

Hence

$$\lim_{n\longrightarrow\infty}\sup_{\delta\to 0}\left|\frac{T_n(\ \theta^*)_{i,j}}{n\delta}\right| \leq \lim_{n\longrightarrow\infty}\|\frac{1}{n}\nabla\frac{\partial^2 Q_n}{\partial\theta_i\partial\theta_j}(\ \theta_0)\| + \sup_{\delta\to 0}|R(\ \theta^* - \theta_0)|.$$

Finally, if we consider δ^* small enough such that $\|\theta^* - \theta_0\| < \delta^* < \delta$, it follows that $R \to 0$. By **A.**5.3 we have the boundedness of the third partial derivative, and applying the Ergodic Theorem, together with the moment assumptions we can conclude the proof. ∎

Theorem 5.6 *Let us assume **A.**5.1-**A.**5.3 to hold, $\varepsilon > 0, \delta > 0$ given and $\mathcal{B}(\ \theta_0, \delta)$ the open sphere of radius δ centered at θ_0. Then, for some $\delta^*, 0 < \delta^* < \delta$, there exists E with $\mathbb{P}(\ E) > 1 - \varepsilon$ and n_0 such that on E for any $n > n_0$, the least squares equations have the solution $\{\theta_n\}$ in $\mathcal{B}(\ \theta_0, \delta)$ at which point $Q_n(\ \theta)$ attains a relative minimum.*

This theorem implies that the solutions of the least squares problem we obtain under our assumptions are local minima.

Proof: For the proof we need to observe that this theorem is a version of Theorem

2.1 in Klimko and Nelson [55] for i.i.d. random variables. Additionally, Proposition 5.4 supplies the necessary assumptions. ∎

Theorem 5.7 *(Strong consistency under weak conditions)*
Under the assumption of Theorem 5.6, there exists $\{\theta_n\}$ *such that* $\theta_n \to \theta_0$ *a.s., and for* $\varepsilon > 0$*, there is* E *with* $\mathbb{P}(\ E) > 1 - \varepsilon$ *and* n_0 *such that on* E *for* $n > n_0, \theta_n$ *satisfies the least squares equation and* Q_n *attains a relative minimum at* θ_n

Proof: This is a direct corollary of the previous theorem and the proof follows just as in Corollary 2.1 in Klimko and Nelson [55]. ∎

The last theorem just proves the existence of a sequence of minimizers which converges almost surely toward the predefined optimal network weight vector. But the theorem does not tell us how to find this optimal parameter vector. The chapter on Backpropagation will provide a numerical approach for finding the solution to this problem.

5.1.3 Asymptotic Normality

To state the asymptotic normality of such a consistent sequence we need some additional assumptions on the mixing coefficients of X_t, i.e. their decreasing rate and some strong regularity considerations on the activation function (or more generally on the network function).

A. 5.8

$$\mathbb{E}\,|X_t|^{2\gamma} < \infty \text{ for some } \gamma > 2$$
$$\alpha(\ k) \leq qk^{-\beta} \text{ for some } q > 0 \quad and \quad \beta > \frac{\gamma}{\gamma - 1}$$

A. 5.9 $f(x, \theta)$*, its first and second partial derivatives (w.r.t* θ *) are measurable with respect to* x *and uniformly continuous*[2] *in a neighborhood of* θ_0 *for every* x*.*

Given these assumptions, let us move toward the statement and the proof of the asymptotic normality.

Proposition 5.10 *Let us assume* **A.***5.1 to* **A.***5.8 hold, then*

$$n^{-1/2}\nabla Q_n(\ \theta_0) \longrightarrow \mathcal{N}\ (\ 0, B(\ \theta_0))\ \ as\ n \longrightarrow \infty$$

[2]Consider $A \subset X$, X and Y are metric spaces. $f : A \longrightarrow Y$ is uniformly continuous on A if $\forall\ \epsilon \quad \exists\ \delta \quad \forall\ x, y \in A, \quad \|x - y\|_X < \delta \Longrightarrow \|f(\ x) - f(\ y)\|_Y < \epsilon$
Uniformly continuous functions have the property to map Cauchy sequence to Cauchy sequence and therefore preserve the uniform convergence of sequence of functions.

where

$$B(\theta_0) = \int (X_t - f(\mathbb{X}_{t-1}, \theta_0))^2 \nabla f(\mathbb{X}_{t-1}, \theta_0) \nabla f(\mathbb{X}_{t-1}, \theta_0)' d\mathbb{P}$$

This proposition helps us to formulate and prove the final result on the asymptotic normality. Before that proof, we need some auxiliary results, for example the following lemma.

Lemma 5.11 *Let X_t be a strictly stationary process and α-mixing; with appropriate decreasing coefficients. Let f be a measurable function from $\mathbb{R}^p$ to $\mathbb{R}$. If we define*

$$Y_t = f(X_{t-1}, \cdots, X_{t-p}),$$

then Y_t is strictly stationary and satisfies an α-mixing condition with decreasing mixing coefficient of the same order as that of X_t

This lemma has been explicitly formulated and prove in Chapter 3 and to improve the readability we also present the proof in this context.

Proof: (Proof of Lemma 5.11)

Obviously X_t is strictly stationary. Let us recall that

$$\alpha(k) = \sup_{A \in \mathcal{F}_{-\infty}^n, B \in \mathcal{F}_{n+k}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

Observing that

$$\mathcal{F}_{-\infty}^0 = \sigma(Y_{-\infty}, \cdots, Y_0) \subseteq \sigma(X_{-\infty}, \cdots, X_{-1}) = \mathcal{G}_{-\infty}^{-1}$$
$$\mathcal{F}_k^\infty = \sigma(Y_k, \cdots, Y_\infty) \subseteq \sigma(X_{k-p}, \cdots, X_\infty) = \mathcal{G}_{k-p}^\infty$$

we have the following:

$$\begin{aligned} \alpha_Y(k) &= \sup_{A \in \mathcal{F}_{-\infty}^n, B \in \mathcal{F}_{n+k}^\infty} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| \\ &\leq \sup_{\tilde{A} \in \mathcal{G}_{-\infty}^{n-1}, \tilde{B} \in \mathcal{G}_{n+k-p}^\infty} \left|\mathbb{P}(\tilde{A} \cap \tilde{B}) - \mathbb{P}(\tilde{A})\mathbb{P}(\tilde{B})\right| = \alpha_X(k-p+1), \end{aligned}$$

if we consider $k > p - 1$. ∎

Now, let us focus on the proof of proposition 5.10.

Proof: To prove this proposition we need to use the Cramer-Wold Device (compare Billingsley [9] page 49 and Proposition 1.6.5 in Brockwell and Davis [12]), i.e. it suffices to show that each linear combination of the components of $n^{-1/2}\nabla Q_n(\theta_0)$

converges in distribution to a suitable linear combination of univariate normal distributions. By definition, we have

$$\begin{aligned}\sum_{i=1}^{l}\lambda_i\frac{\partial Q_n}{\partial\theta_i}(\ \theta_0) &= \sum_{i=1}^{l}\lambda_i\left(\sum_{t=1}^{n}\left((\ X_t - f(\ \mathbb{X}_{t-1},\theta_0))\frac{\partial f(\ \mathbb{X}_{t-1},\theta_0)}{\partial\theta_i}\right)\right)\\ &= \sum_{i=1}^{l}\lambda_i\left(\sum_{t=1}^{n}X_{t,i}\right)\\ &= \sum_{i=1}^{l}\lambda_i S_{n,i}.\end{aligned}$$

Without loss of generality, we will assume $\{X_{t,i}\}$ to be a zero mean process, otherwise we just need to use the shifted versions of these processes.
Let us observe that $\{X_{t,i}\}$ inherits some of the properties of the process $\{X_t\}$ for example, stationarity and mixing property. Moreover, the mixing coefficients of $\{X_{t,i}\}$ have the same decreasing rate (up to a constant factor) as those of $\{X_t\}$ as proved in Lemma 5.11 From the last observation one can use a Central Limit Theorem for an α-mixing stationary process (see e.g., Bosq [10] Theorem 1.7) together with Slutsky's Lemma to conclude the proof. ∎

Given the proof of this proposition, we have almost all the ingredients to state and give a proof of the asymptotic normality of the parameter estimates. Let us first present an intermediate step.

Lemma 5.12 *(Ergodic Lemma for a Triangular Array)*
Let X_t be a strictly stationary and ergodic process on a given probability space with values in $\mathbb{R}^u$ and let Z_n be a sequence of $\mathbb{R}^k$-valued random variables such that

$$\lim_{n\to\infty} Z_n = Z_\infty\ a.s. \tag{5.9}$$

Let $g(\ y,z)$ be a mapping from $\mathbb{R}^{u+k}$ to $\mathbb{R}$ which is Borel-measurable in y and uniformly continuous in z for every y and such that $\mathbb{E}|g(\ X_1,Z_\infty)|<\infty$. Then,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}g(\ X_t,Z_n) = \mathbb{E}g(\ X_1,Z_\infty). \tag{5.10}$$

Proof: See Dimitroff [15] ∎

The following theorem yields the result on the asymptotic normality of the parameter estimates.

Theorem 5.13 *Let us assume **A**.5.1 to **A**.5.9 hold, then*

$$n^{1/2}(\ \theta_n-\theta_0)\to\mathcal{N}(\ 0,\Sigma(\ \theta_0))\ \textit{as}\ n\longrightarrow\infty,$$

where

$$\Sigma = A^{-1}(\theta_0)B(\theta_0)A^{-1}(\theta_0), \quad A = \left(\mathbb{E}\left(\frac{\partial^2 (X_t - f(\mathbb{X}_{t-1}, \theta_0))^2}{\partial\theta_i \partial\theta_j} \right) \right)_{i,j}$$

$$B(\theta_0) = \int (X_t - f(\mathbb{X}_{t-1}, \theta_0))^2 \nabla f(\mathbb{X}_{t-1}, \theta_0) \nabla f(\mathbb{X}_{t-1}, \theta_0)' d\mathbb{P}.$$

As in linear regression, the practical importance of this theorem is that one can use it to set confidence intervals for the unknown parameter. In the Neural Network framework, one can use it to test whether all the parameter of the network are relevant.

Proof: Let us assume that θ_n satisfies the least squares equations. Using a Taylor expansion around θ_0 for $n^{-1/2}\nabla Q_n$ one can write

$$\begin{aligned} 0 &= n^{-1/2}\nabla Q_n(\theta_n) \\ &= n^{-1/2}\nabla Q_n(\theta_0) + n^{-1}\left(\frac{\partial^2 Q_n}{\partial\theta_i \partial\theta_j}(\theta_n^*) \right)_{i,j} n^{1/2}(\theta_n - \theta_0), \end{aligned}$$

with $\theta_n^* = \theta_0 + \lambda(\theta_n - \theta_0)$ for some $\lambda \in (0, 1)$.
From the previous theorem we have that $\theta_n^* \to \theta_0$ a.s. and by proposition 5.10, $n^{-1/2}\nabla Q_n(\theta_0)$ is asymptotically normal . Therefore, we only need to investigate the asymptotic behavior of $n^{-1}\left(\frac{\partial^2 Q_n}{\partial\theta_i \partial\theta_j}(\theta_n^*) \right)_{i,j}$. Using **A.**5.9, together with the Ergodic Lemma for a Triangular Array, one has

$$n^{-1}\left(\frac{\partial^2 Q_n}{\partial\theta_i \partial\theta_j}(\theta_n^*) \right)_{i,j} \to A(\theta_0)$$

and the proof follows. ∎

In this section we have established the consistency and the asymptotic normality of the parameter estimates for the nonlinear least squares. In the coming section we propose an extension of these asymptotic properties to the parameter estimates of a weighted nonlinear least squares as it will be defined.

5.2 Nonlinear Weighted Least Squares

In this section we consider a training set of the random variable

$$X_t = \sum_{k=1}^{K} S_{t,k} X_{t,k} \text{ with } S_{t,k} = \begin{cases} 1 & \text{for } Q_t = k \\ 0 & \forall\, Q_t \neq k \end{cases} \tag{5.11}$$

$$t = -\tau + 1, \cdots, -p + 1, \cdots n, \cdots, n + \tau + 1, \tag{5.12}$$

where τ is a given **weighting** parameter that we will precise later on and p the common order of the underlying $NLAR$ processes $X_{t,k} = m_k(\ X_{t-1}, \cdots, X_{t-p}) + Z_{t,k}$. If the parameters $p(\ k)$ are not equal, we simply choose the maximum of these parameters and set it to p. Remark that this type of mixture model is different to the one we consider in the previous chapters. There, the dynamic of the time series changes sometimes, but here, there are K stationary background processes $X_{t,1}, \cdots, X_{t,K}$ and the state variable Q_t only determine which time series we observe at time t. Q_t has no longer to be a Markov Chain, but an arbitrary stationary process independent of the $X_{t,k}, k = 1, \cdots, K$.
In this context we want to estimate the change-points and the auto-regression function $m_1, \cdots, m_K$ simultaneously. For this purpose, we follow a proposal by Müller et al. [73] and consider $L \geq K$ competing neural networks. For each network $n_l, l \in I_L$ the output function can be written as

$$f_l(\ x_1, \cdots, x_p; \theta^{(\ l)}) = \nu_{0,l} + \sum_{h=1}^{H(\ l)} \nu_h \psi(\ \omega_h^{(\ 0),l} + \sum_{i=1}^{p} \omega_h^{(\ i),l} x_i), l = 1, \cdots, L,$$

where the $\theta^{(\ l)} \in \mathbb{R}^{s(\ l)}(\ $ with $s(\ l) = H(\ l)(\ p + 2) + 1$ are (not necessarily) different. These parameters $\theta^{(\ 1)}, \cdots, \theta^{(\ L)}$ are estimated by the Weighted Nonlinear Least Squares scheme:

$$\begin{aligned} Q_{n,G}(\ \theta^{(\ 1)}, \cdots, \theta^{(\ L)}) &= \sum_{l=1}^{L} \sum_{t=1}^{n} P_t^l e_t^l = \sum_{t=1}^{n} \sum_{l=1}^{L} P_t^l e_t^l \qquad (5.13) \\ &= \sum_{t=1}^{n} q_{t,G} \qquad (5.14) \end{aligned}$$

where

$$e_t^l = (\ X_{t+1} - f_l(\ \mathbb{X}_t, \theta^l))^2 \qquad (5.15)$$

and, using a symmetric moving average of length $2\tau + 1$ around t,

$$P_t^l = \frac{\exp\left(\ -\beta \sum_{j=t-\tau}^{t+\tau} e_n^l\right)}{\sum_{\lambda=1}^{L} \exp\left(\ -\beta \sum_{j=t-\tau}^{t+\tau} e_j^\lambda\right)}. \qquad (5.16)$$

P_t^l can be interpreted as the current (in term of the estimation algorithm) estimate of the probability that $Q_t = l$, i.e. the probability that we observe the time series with autoregressive function approximated by the l'th network. If, around time t, the network function f_l fits the data well, then the e_j^l will be small for $|t - j| \leq \tau$, and P_t^l will be large. To achieve it, we implicitly have to solve the minimization problem

$$\min_{\theta_G} \mathbb{E} q_{t,G}(\ \theta^{(\ 1)}, \cdots, \theta^{(\ L)}) = \min_{\theta_G} Q^G(\theta_G), \qquad (5.17)$$

where $\theta_G = \left(\theta^{(1)}, \cdots, \theta^{(L)}\right)$, i.e. we need to minimize the expectation of the weighted squared errors.

Remark 5.14 *At this point, let us note that we have a function that is minimize and for which we can explore the asymptotic property of the parameter estimate. This is the case in the approach considered by Müller et al in [73].*

Before we go through the resolution of this problem, let us first make some comments on the weighting parameter τ and the parameter β.

Remark 5.15 *As far as τ is concerned, this parameter can be regarded in our case as the bandwidth in the Kernel type estimates, in the sense that if we take τ to be small, very few data will contribute to the estimation of the changepoint and conversely a large amount of data will contribute to determination of the changepoint if we consider τ to be large enough. If τ is small, we will detect a change in Q_t fast, but not reliably, whereas for τ large changes will be detected and we have no false alarms with high probability, but it will take a lot of time. Also, short periods of observing a different time series may be overlooked.*
Unlike the inverse temperature in the Simulated Annealing, as β will increase, if we take the other value to be constant, the networks producing the largest errors around the changepoint will produce weights that should fall quickly under a given threshold and therefore will have little effect on the dynamic of the process around this point. Additionally we need to recall that β has to be a positive real quantity for which we need to find the suitable value. In this light we follow a proposal by Kroisandt, i.e. for the numerical procedure we will consider an increasing real sequence for β. Nevertheless we need to start with a value of β that is strictly positive in order to prevent the entire system to collapse, i.e we want the entire system to satisfy some regularity condition all over the computation steps.

Once our goal for this section is clearly defined, we can move forward in its resolution and once more we have to find the asymptotic properties of the parameter estimates. Let us first present some preliminary results.

5.2.1 Preliminaries

Recall that

$$Q_{n,G}(\theta^{(1)}, \cdots, \theta^{(L)}) = \sum_{l=1}^{L}\sum_{t=1}^{n} P_t^l e_t^l = \sum_{t=1}^{n}(\sum_{l=1}^{L} P_t^l e_t^l) \tag{5.18}$$

$$= \sum_{t=1}^{n} q_{t,G} \tag{5.19}$$

with

$$e_t^l = (\ X_{t+1} - f(\ \mathbb{X}_t, \theta^l))^2 \tag{5.20}$$

and

$$P_t^l = \frac{\exp\left(-\beta \sum_{j=t-\tau}^{t+\tau} e_j^l\right)}{\sum_{\lambda=1}^{L} \exp\left(-\beta \sum_{j=t-\tau}^{t+\tau} e_j^\lambda\right)}. \tag{5.21}$$

To improve the readability, let us set

$$g(\ t, \lambda) = -\beta \sum_{j=t-\tau}^{t+\tau} e_j^\lambda.$$

For the computation of the partial derivatives, we need

$$\frac{\partial e_t^l}{\partial \theta_i} = \begin{cases} 0 & \text{if } \theta_i = \theta_i^\lambda \text{ for } l \neq \lambda \\ \frac{\partial f}{\partial \theta_i^l}(\ \mathbb{X}_t, \theta^l)(\ X_{t+1} - f(\ \mathbb{X}_t, \theta^l)) & \text{if } \theta_i = \theta_i^l \end{cases} \tag{5.22}$$

and

$$\frac{\partial P_t^l}{\partial \theta_i} = \begin{cases} P_t^l(\ 1 - P_t^l)\frac{\partial g(\ t,l)}{\partial \theta_i^l} & \text{if } \theta_i = \theta_i^l \\ -P_t^l P_t^\lambda \frac{\partial g(\ t,\lambda)}{\partial \theta_i^\lambda} & \text{if } \theta_i = \theta_i^\lambda \text{ for } l \neq \lambda. \end{cases} \tag{5.23}$$

Hence,

$$\begin{aligned} & \frac{\partial q_{t,G}}{\partial \theta_i^\lambda} \\ = & \sum_{l=1}^{L} \frac{\partial(\ P_t^l e_t^l)}{\partial \theta_i^\lambda} \\ = & -\sum_{l \neq \lambda} P_t^l P_t^\lambda e_t^l \frac{\partial g(\ t,\lambda)}{\partial \theta_i^\lambda} + P_t^\lambda(\ 1 - P_t^\lambda) e_t^\lambda \frac{\partial g(\ t,\lambda)}{\partial \theta_i^\lambda} + P_t^\lambda \frac{\partial e_t^\lambda}{\partial \theta_i^\lambda} \\ = & -q_{t,G} P_t^\lambda \frac{\partial g(\ t,\lambda)}{\partial \theta_i^\lambda} + P_t^\lambda e_t^\lambda \frac{\partial g(\ t,\lambda)}{\partial \theta_i^\lambda} + P_t^\lambda \frac{\partial e_t^\lambda}{\partial \theta_i^\lambda}. \end{aligned}$$

From this we derive for the extreme cases, i.e. the situation where higher moment assumptions should be needed

$$\frac{\partial^2 q_{t,G}}{\partial \theta_i^\lambda \partial \theta_j^\nu} = -P_t^\nu P_t^\lambda \left(e_t^\nu \frac{\partial g(\ t,\nu)}{\partial \theta_j^\nu} + \frac{\partial e_t^\nu}{\partial \theta_j^\nu} \right) \frac{\partial g(\ t,\lambda)}{\partial \theta_i^\lambda},$$

$$\begin{aligned}
\frac{\partial^2 q_{t,G}}{\partial\theta_i^\lambda\partial\theta_j^\lambda} &= q_{t,G}P_t^\lambda\left(\,(\,2P_t^\lambda-1)\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}+\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_j^\lambda}\right)\\
&+ P_t^\lambda(\,1-P_t^\lambda)\left(\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial e_t^\lambda}{\partial\theta_j^\lambda}+\frac{\partial e_t^\lambda}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}\right)\\
&+ P_t^\lambda e_t^\lambda\left(\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_j^\lambda}+(\,1-2P_t^\lambda)\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}\right)\\
&+ P_t^\lambda\frac{\partial^2 e_t^\lambda}{\partial\theta_i^\lambda\partial\theta_j^\lambda}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial^3 q_{t,G}}{\partial\theta_i^\lambda\partial\theta_j^\lambda\partial\theta_k^\lambda}\\
&= q_{t,G}P_t^\lambda[(\,6P_t^\lambda(\,1-P_t^\lambda)-1)\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_k^\lambda}\\
&+ (\,2P_t^\lambda-1)\left(\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_k^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}+\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_j^\lambda\partial\theta_k^\lambda}\right)]\\
&+ P_t^\lambda P_t^\lambda\left(e_t^\lambda\frac{\partial g(\,t,\lambda)}{\partial\theta_k^\lambda}+\frac{\partial g(\,t,\lambda)}{\partial\theta_k^\lambda}\right)\left((\,2P_t^\lambda-1)\left(\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}+\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_j^\lambda}\right)\right)\\
&+ P_t^\lambda(\,1-3P_t^\lambda-2P_t^\lambda P_t^\lambda)\frac{\partial g(\,t,\lambda)}{\partial\theta_k^\lambda}\left(\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial e_t^\lambda}{\partial\theta_j^\lambda}+\frac{\partial e_t^\lambda}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}\right)\\
&+ P_t^\lambda(\,1-P_t^\lambda)\left(\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_k^\lambda}\frac{\partial e_t^\lambda}{\partial\theta_j^\lambda}+\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial^2 e_t^\lambda}{\partial\theta_j^\lambda\partial\theta_k^\lambda}+\frac{\partial^2 e_t^\lambda}{\partial\theta_i^\lambda\partial\theta_k^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}+\frac{\partial e_t^\lambda}{\partial\theta_i^\lambda}\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_j^\lambda\partial\theta_k^\lambda}\right)\\
&+ P_t^\lambda(\,1-P_t^\lambda)e_t^\lambda\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\left(\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_j^\lambda}+(\,1-2P_t^\lambda)\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}\right)\\
&+ P_t^\lambda e_t^\lambda\left(\frac{\partial^3 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_j^\lambda\partial\theta_k^\lambda}-2P_t^\lambda(\,1-P_t^\lambda)\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_k^\lambda}\right)\\
&+ (\,1-2P_t^\lambda)P_t^\lambda e_t^\lambda\left(\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_i^\lambda\partial\theta_k^\lambda}\frac{\partial g(\,t,\lambda)}{\partial\theta_j^\lambda}+\frac{\partial g(\,t,\lambda)}{\partial\theta_i^\lambda}+\frac{\partial^2 g(\,t,\lambda)}{\partial\theta_j^\lambda\partial\theta_k^\lambda}\right)\\
&+ P_t^\lambda(\,1-P_t^\lambda)\frac{\partial g(\,t,\lambda)}{\partial\theta_k^\lambda}\frac{\partial^2 e_t^\lambda}{\partial\theta_i^\lambda\partial\theta_j^\lambda}+P_t^\lambda\frac{\partial^3 e_t^\lambda}{\partial\theta_i^\lambda\partial\theta_j^\lambda\theta_k^\lambda}.
\end{aligned}$$

5.2.2 Consistency

As for the Nonlinear Least Squares problem, in order to solve this problem, we essentially want to find the solution of the "Weighted Least Squares Equations":

$$\frac{\partial Q_{n,G}}{\partial \theta_i} = 0 \ \forall i = 1, \cdots, s; s = \sum_{l=1}^{L} s(\ l). \tag{5.24}$$

As previously, we would like to control the behavior of the Taylor expansion of $Q_{n,G}$ around the optimal parameter given it exists. To do so, we have to state some sets of assumptions which are in a certain sense much stronger than what we have stated for the least squares problem. This could be justifying why we started from the least squares problem, although one can consider the results for the least squares problem as corollaries of the much more general results which we are going to present in the coming lines.

A. 5.16 *1. Assume that* $\tau \in \mathbb{N}$ *and* $p \in \mathbb{N}; 1 \leq p \leq \tau$.

2. $S_{t,k}$ *is strictly stationary,* $\alpha-$*mixing and independent of* $X_{t,k}$, *for each* $k = 1, \cdots, K$.

A. 5.17 *1.* $Z_{t,k},\ k = 1, \cdots, K$ *i.i.d. with mean* 0.

2. $\exists\ C_j > 0$ *such that* $\sup_{k \in I_K} \mathbb{E}(\ |Z_{t,k}|^j \mid \mathbb{X}_t = x) \leq C_j < \infty$ *for* $j = 1, \cdots, 9$.

3. $m_k(x), k = 1, \cdots, K$ *are continuous and* $\exists \kappa \geq 0, a_i \geq 0$ *for* $i = 1 \cdots, p$ *such that* $\sup_{k \in I_K} |m_k(x)| \leq \sum_{i=1}^{p} a_i |x_i| + \kappa$.

A. 5.18 *1.* $X_{t,k}, k = 1, \cdots, K$ *are independent NLAR (p)-processes;*

2. $X_{t,k}$ *is strictly stationary and* $\alpha-$*mixing for each* $k \in I_K$.

3. $\mathbb{E}|X_t|^{4\gamma+1} < \infty$ *for some* $\gamma > 2$.

Remark 5.19 *1. From the fact that* $X_{t,k}, k \in I_K$ *are independent, strictly stationary stochastic processes and* $S_{t,k}$ *is strictly stationary and independent of* $X_{t,k}$ *for each* $k = 1, \cdots, K$, *it follows that* X_t *is strictly stationary.*

2. X_t *satisfies an* $\alpha-$*mixing condition with mixing coefficients for which the decreasing rate is determined by the slowest decreasing rate of the mixing coefficients of* $X_{t,k}, k \in I_K$ *and of* $S_{t,k}$.

The above remark is formalized in the following lemma.

Lemma 5.20 *Let $\{X_{t,k}\}, k \in I_K$ be strictly stationary and independent stochastic processes, each satisfying an $\alpha-$mixing condition. Also consider $\{S_{t,k}\}$ to be a strictly stationary stochastic process independent of $X_{t,k}$ for each $k \in I_K$. Moreover, assume $S_{t,k}$ satisfies an $\alpha-$mixing condition. Assume that $\rho_t \to 0$ denotes a sequence of upper bounds for the mixing coefficients of $X_{t,k}, k \in I_K, S_t$. Then $\{X_t\}$ is strictly stationary and satisfies an $\alpha-$mixing condition with coefficient coefficients bounded by $(K+1)\rho_t$.*

Proof: It is that X_t is strictly stationary, let us concentrate on the proof of the second conclusion of the lemma. For that purpose, let us define the following:

$$\mathcal{F}_0^{(\ S)} = \sigma(\ S_{t,k}, t \leq 0) \ \ \mathcal{F}_{u+}^{(\ S)} = \sigma(\ S_{t,k}, t \geq u)$$

Analogously we define
$\mathcal{F}_0^{(\ X_k)}$, $\mathcal{F}_{u+}^{(\ X_k)}$, for each k and $\mathcal{F}_0^{(\ X)}, \mathcal{F}_{u+}^{(\ X)}$. It follows that

$$\mathcal{F}_0^{(\ X)} \subset \left(\underset{k}{\cup} \mathcal{F}_0^{(\ X_k)} \right) \cup \mathcal{F}_0^{(\ S)}$$

and

$$\mathcal{F}_{u+}^{(\ X)} \subset \left(\underset{k}{\cup} \mathcal{F}_{u+}^{(\ X_k)} \right) \cup \mathcal{F}_{u+}^{(\ S)}.$$

Using Theorem 1 p 4 from Doukhan [18], we can conclude that

$$\alpha(\ \mathcal{F}_0^{(\ Y)}, \mathcal{F}_{u+}^{(\ Y)}) \leq \alpha(\ \mathcal{F}_0^{(\ S)}, \mathcal{F}_{u+}^{(\ S)}) + \sum_{k=1}^{K} \alpha(\ \mathcal{F}_0^{(\ X_k)}, \mathcal{F}_{u+}^{(X_k)})$$

from which it follows that X_t is $\alpha-$mixing. Using Lemma $B.2.2$ page 175 in Kroisandt [57], we conclude with the rate of convergence. ∎

Remark 5.21 *By means of the Geometric Ergodic Theory we have established the asymptotic stationarity of $\{X_t\}$ under weak assumptions. Hence all the assumption of the previous lemma can be considered as technical tools to achieve the results on the asymptotic properties of the parameter estimates in the weighted least squares approach.*

A. 5.22 *For each network $n_l, l = 1, \cdots, L$ the activation functions ψ are bounded $\mathcal{C}^3$ with bounded derivatives.*

A. 5.23 *Assume that Q^G attains it unique global minimum $(\theta_0^{(1)}, \cdots, \theta_0^{(L)})$ which is an interior point of Θ_G (compact subspace of $\mathbb{R}^s$) and also assume $\nabla^2 Q^G(\theta_0^{(1)}, \cdots, \theta_0^{(L)})$ is positive definite.*

Lemma 5.24 *Consider assumptions* **A.***5.16-***A.***5.22 to hold. Then, it follows that*

$$\mathbb{E}\left|e_t^l(\theta_0^{(1)},\cdots,\theta_0^{(L)})\right|^4 < \infty \; \forall l = 1,\cdots,L; \tag{5.25}$$

$$\mathbb{E}\left|\frac{\partial g(t,l)}{\partial \theta_i}(\theta_0^{(1)},\cdots,\theta_0^{(L)})\right|^4 < \infty \; \forall \; i = 1,\cdots,s \tag{5.26}$$

$$\mathbb{E}\left|\frac{\partial^2 g(t,l)}{\partial \theta_i \partial \theta_j}(\theta_0^{(1)},\cdots,\theta_0^{(L)})\right|^3 < \infty \; \forall \; i,j = 1,\cdots,s \tag{5.27}$$

$$\mathbb{E}\left|\frac{\partial^3 g(t,l)}{\partial \theta_i \partial \theta_j \partial \theta_k}(\theta_0^{(1)},\cdots,\theta_0^{(L)})\right|^2 < \infty \; \forall \; i,j,k = 1,\cdots s. \tag{5.28}$$

Proof: By definition and using the triangular inequality together with the fact that $S_{t,k}$ is bounded by 1 we have

$$\begin{aligned}
\left|X_{t+1} - f(\mathbb{X}_t,\theta^l)\right| &= \left|\sum_{k=1}^{K} S_{t,k}(X_{t+1} - m_k(\mathbb{X}_t)) \right. \\
&\quad \left. + \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_t)) - f(\mathbb{X}_t,\theta^l))\right| \\
&\leq \sum_{k=1}^{K} |X_{t+1} - m_k(\mathbb{X}_t)| + \sum_{k=1}^{K} \left|S_{t,k}(m_k(\mathbb{X}_t) - f(\mathbb{X}_t,\theta^l)\right|.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}(e_t^l)^4 &= \mathbb{E}\left|X_{t+1} - f(\mathbb{X}_t,\theta^l)\right|^8 \\
&\leq \mathbb{E}\left(\sum_{k=1}^{K} |X_{t+1} - m_k(\mathbb{X}_t)| + \sum_{k=1}^{K} |S_{t,k}(m_k(\mathbb{X}_t) - f(\mathbb{X}_t,\theta^l))|\right)^8 \\
&\leq const.\left(\sum_{k=1}^{K} \mathbb{E}|X_{t+1} - m_k(\mathbb{X}_t)|^8 \right. \\
&\quad \left. + \mathbb{E}\sum_{k=1}^{K} |S_{t,k}(m_k(\mathbb{X}_t) - f(\mathbb{X}_t,\theta^l))|^8\right),
\end{aligned}$$

by using the fact that for $a,b,\tau \geq 0, (a+b)^\tau \leq 2^\tau(a^\tau + b^\tau)$ and the finiteness of $\mathbb{E}|e_t^l|^4$ it follows from assumptions **A.**5.16 to **A.**5.22.
With a similar strategy we will find the necessary conditions to achieve the remaining part of the proof .
By definition, we have

$$\frac{\partial g(t,l)}{\partial \theta_i} = \beta \sum_{j=t-\tau}^{t+\tau} \frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)(X_{j+1} - f(\mathbb{X}_j,\theta^l)). \tag{5.29}$$

Therefore,

$$\begin{aligned}\left|\frac{\partial g(t,l)}{\partial \theta_i}\right| &= \left|\beta \sum_{j=t-\tau}^{t+\tau} \frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right| \\ &\leq \beta \sum_{j=t-\tau}^{t+\tau} \left|\frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)||X_{j+1}-f(\mathbb{X}_j,\theta^l\right|.\end{aligned}$$

Thus,

$$\begin{aligned}\left|\frac{\partial g(t,l)}{\partial \theta_i}\right|^4 &= \left|\beta \sum_{j=t-\tau}^{t+\tau} \frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^4 \\ &\leq Const. \sum_{j=t-\tau}^{t+\tau} \left|\frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^4 \\ &\leq Const. \left(\sum_{j=t-\tau}^{t+\tau} \left|\frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^8 \right. \\ &\quad \left. + \sum_{j=t-\tau}^{t+\tau} \left|X_{j+1}-f(\mathbb{X}_j,\theta^l)\right|^8 \right)\end{aligned}$$

and the finiteness of $\mathbb{E}|\frac{\partial g(t,l)}{\partial \theta_i}|^4$ is easily provided by our assumptions.
Analogous to the previous two definitions we have

$$\begin{aligned}|\frac{\partial^2 g(t,l)}{\partial \theta_i \partial \theta_j}| &\leq \beta \sum_{j=t-\tau}^{t+\tau} \left(\left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right| \right. \\ &\quad \left. + \left|\frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial \theta_j}(\mathbb{X}_j,\theta^l)\right| \right),\end{aligned}$$

from which it follows that

$$\begin{aligned}\left|\frac{\partial^2 g(t,l)}{\partial \theta_i \partial \theta_j}\right|^3 &\leq Const.\beta \sum_{j=t-\tau}^{t+\tau} \left(\left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^3 \right. \\ &\quad \left. + \left|\frac{\partial f}{\partial \theta_i}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial \theta_j}(\mathbb{X}_j,\theta^l)\right|^3 \right).\end{aligned}$$

By an application of a Hölder type inequality it follows that

$$\begin{aligned}&\mathbb{E}\left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^3 \\ =\ &\mathbb{E}\left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\mathbb{X}_j,\theta^l)|^3|X_{j+1}-f(\mathbb{X}_j,\theta^l)\right|^3 \\ \leq\ &\left(\mathbb{E}\left|\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}(\mathbb{X}_j,\theta^l)\right|^{9/2}\right)^{2/3} \left(\mathbb{E}\left|X_{j+1}-f(\mathbb{X}_j,\theta^l)\right|^9\right)^{1/3},\end{aligned}$$

from which we can deduce the finiteness of

$$\left|\frac{\partial^2 g(t,l)}{\partial\theta_i\partial\theta_j}\right|^3 .$$

In the final step of the proof of this lemma, we focus on

$$\begin{aligned}\left|\frac{\partial^3 g(t,l)}{\partial\theta_i\partial\theta_j\theta_k}\right| &= \Bigg|\beta\sum_{j=t-\tau}^{t+\tau}\Bigg(\frac{\partial^3 f}{\partial\theta_i\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\\ &\quad -\frac{\partial^2 f}{\partial\theta_i\partial\theta_j}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_k}(\mathbb{X}_j,\theta^l)\\ &\quad -\frac{\partial^2 f}{\partial\theta_i\partial\theta_k}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_j}(\mathbb{X}_j,\theta^l)\\ &\quad -\frac{\partial^2 f}{\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_i}(\mathbb{X}_j,\theta^l)\Bigg)\Bigg|.\end{aligned}$$

Hence, by similar arguments as those previously used,

$$\begin{aligned}\left|\frac{\partial^3 g(t,l)}{\partial\theta_i\partial\theta_j\theta_k}\right|^2 &\le const.\sum_{j=t-\tau}^{t+\tau}\Bigg(\left|\frac{\partial^3 f}{\partial\theta_i\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^2\\ &\quad +\left|\frac{\partial^2 f}{\partial\theta_i\partial\theta_j}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_k}(\mathbb{X}_j,\theta^l)\right|^2\\ &\quad +\left|\frac{\partial^2 f}{\partial\theta_i\partial\theta_k}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_j}(\mathbb{X}_j,\theta^l)\right|^2\\ &\quad +\left|\frac{\partial^2 f}{\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_i}(\mathbb{X}_j,\theta^l)\right|^2\Bigg)\end{aligned}$$

Consequently

$$\begin{aligned}&\mathbb{E}\left|\frac{\partial^3 g(t,l)}{\partial\theta_i\partial\theta_j\theta_k}\right|^2\\ \le\ & const.\sum_{j=t-\tau}^{t+\tau}\Bigg(\mathbb{E}\left|\frac{\partial^3 f}{\partial\theta_i\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^2\\ &+\mathbb{E}\left|\frac{\partial^2 f}{\partial\theta_i\partial\theta_j}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_k}(\mathbb{X}_j,\theta^l)\right|^2\\ &+\mathbb{E}\left|\frac{\partial^2 f}{\partial\theta_i\partial\theta_k}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_j}(\mathbb{X}_j,\theta^l)\right|^2\\ &+\mathbb{E}\left|\frac{\partial^2 f}{\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)\frac{\partial f}{\partial\theta_i}(\mathbb{X}_j,\theta^l)\right|^2\Bigg)\end{aligned}$$

and once more, using a Hölder type inequality

$$\begin{aligned} & \mathbb{E}\left|\frac{\partial^3 f}{\partial\theta_i\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)(X_{j+1}-f(\mathbb{X}_j,\theta^l))\right|^2 \\ \leq & \left(\mathbb{E}\left|\frac{\partial^3 f}{\partial\theta_i\partial\theta_j\partial\theta_k}(\mathbb{X}_j,\theta^l)\right|^3\right)^{2/3}\left(\mathbb{E}\left|X_{j+1}-f(\mathbb{X}_j,\theta^l)\right|^6\right)^{1/3} \end{aligned}$$

The proof follows as in the other cases. ∎

In fact the above lemma helps us to see that we really need the moment assumption we have stated. Indeed it provides a helpful (as we will observe in the proof of the next proposition) set of sufficient conditions which are satisfied under our assumptions.
In the following Proposition, $V_{n,G}$ and $T_{n,G}$ are defined similarly to V_n and T_n, respectively. Moreover we will assume $(\theta_*^{(1)},\cdots,\theta_*^{(L)})$ to be in a suitable neighborhood of $(\theta_o^{(1)},\cdots,\theta_o^{(L)})$.

Proposition 5.25 *Let us assume the assumptions* ***A****.5.16 to* ***A****.5.23 hold, then*

1.
$$\lim_{n\longrightarrow\infty}\frac{1}{n}\frac{\partial Q_{n,G}}{\partial\theta_i}(\theta_o^{(1)},\cdots,\theta_o^{(L)})=0\ a.s.\ \forall\ i=1,\cdots,s$$

2.
$$\lim_{n\longrightarrow\infty}\frac{1}{n}V_{n,G}=A_G(\theta_o^{(1)},\cdots,\theta_o^{(L)})\ \textit{a.s.}$$

3.
$$\lim_{n\longrightarrow\infty}\sup_{\delta\to 0}\left(\frac{|T_{n,G}(\theta_*^{(1)},\cdots,\theta_*^{(L)})_{i,j}|}{n\delta}\right)<\infty\ \textit{a.s. for all}\ i,j=1,\cdots,s.$$

Proof: The proof of this proposition is based on the fact that, from Lemma 5.20, X_t is strictly stationary and satisfies some mixing condition. Having this observation, the rest of the proof follows the proof of Proposition 5.10 word for word. ∎

Having all the above proposition, we have once more stated the conditions of some theorems in [55]. Hence we can easily derive consistency under weak assumptions as we have done in the previous section. We summarize these results in the following theorem.

Theorem 5.26 *(Consistency under Weak Conditions)*
Under assumptions ***A.****5.16 to* ***A.****5.27, there exists* $(\theta_n^{(1)}, \cdots, \theta_n^{(L)})$ *such that*

$$(\theta_n^{(1)}, \cdots, \theta_n^{(L)}) \to (\theta_0^{(1)}, \cdots, \theta_0^{(L)}) a.s.$$

and for $\varepsilon > 0$*, there is* E *with* $\mathbb{P}(E) > 1 - \varepsilon$ *and* n_0 *such that on* E *for* $n > n_0, (\theta_n^{(1)}, \cdots, \theta_n^{(L)})$ *satisfies the weighted least squares equation and* $Q_{n,G}$ ***attains a relative minimum at*** $(\theta_n^{(1)}, \cdots, \theta_n^{(L)})$.

Proof: The proof of this theorem is similar to that of Theorem 5.7 given that all the intermediate steps are achieved with the assumptions stated. ∎

5.2.3 Asymptotic Normality

A. 5.27 $\sup_{k \in I_K} \alpha_{X_k}(q) \leq aq^{-\nu}$ *for some* $a > 0$ *and* $\nu > 1 + \frac{1}{2\gamma - 1}, (I_K = \{1, \cdots, K\})$.

Proposition 5.28 *Let us consider* ***A.****5.16 to* ***A.****5.27 to hold, then*

$$n^{-1/2} \nabla Q_{n,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) \longrightarrow \mathcal{N}\left(0, B_G(\theta_0^{(1)}, \cdots, \theta_0^{(L)})\right),$$

where

$$\begin{aligned} & B_G(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) \\ = & \lim_{n \longrightarrow \infty} \mathbb{E}\, n^{-1} \nabla Q_{n,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) \nabla Q_{n,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)})' \\ = & \mathbb{E} \nabla q_{t,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) \nabla q_{t,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)})'. \end{aligned}$$

Proof: The proof is analogous to the proof of proposition 5.10. ∎

A. 5.29 *For each* $l = 1, \cdots, L,\ f(x, \theta^l)$*, its first and second partial derivatives with respect to* $(\theta^{(1)}, \cdots, \theta^{(L)})$ *are measurable w.r.t* x*, and uniformly continuous in a neighborhood of* $(\theta_0^{(1)}, \cdots, \theta_0^{(L)})$ *for every* x.

Theorem 5.30 *Let us consider* ***A.****5.16 to* ***A.****5.29 hold, then*

$$n^{1/2}((\theta_n^{(1)}, \cdots, \theta_n^{(L)}) - (\theta_0^{(1)}, \cdots, \theta_0^{(L)})) \to \mathcal{N}\left(0, \Sigma_G(\theta_0^{(1)}, \cdots, \theta_0^{(L)})\right),$$

where

$$\begin{aligned} & \Sigma_G = A_G^{-1}(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) B_G(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) A_G^{-1}(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) \\ & A_G(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) = \lim_{n \longrightarrow \infty} \frac{1}{n} V_{n,G} \\ & B_G(\theta_0) = \lim_{n \longrightarrow \infty} \mathbb{E}\, n^{-1} \nabla Q_{n,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)}) \nabla Q_{n,G}(\theta_0^{(1)}, \cdots, \theta_0^{(L)})'. \end{aligned}$$

Proof: This proof follows the proof of Theorem 5.13. ∎

In this section we have studied the consistency and the asymptotic normality of the parameter estimates for the non linear least squares problem and the weighted non linear least squares as well. Back to White [91] we have to observe that the expected consistency here just means convergence toward a local minimum of the criterion function. Moreover there is no guarantee to attain a global minimum. In general, asymptotic normality helps to set down test for the significance of the parameters. However, for the special case of the weighted least squares, it just means that the weights of the non divergent networks can achieve similar test of hypothesis on their significance if the heat parameter is kept constant. Additionally this parameter need to be kept away from zero to make sure that the system satisfies some regularity assumptions.

6 Multivariate Weighted Least Squares for Change-point Analysis in Time Series Models

Unlike the previous chapter, the aim of the current chapter is to extend and prove the key results established for univariate time series to higher dimensional time series. To achieve this goal we need to set some sufficient conditions and this is similar to what we have done in the previous chapter in two stages. In the first stage we will find the asymptotic of the parameter estimates of a nonlinear Multivariate Least Squares and in the second stage extend the results to the nonlinear Multivariate Weighted Least Squares.

6.1 Multivariate Least Squares

Let us assume that we have a nonlinear multivariate time series which can be written in the form

$$X_{t+1} = m(X_t, \cdots, X_{t+1-p}) + \epsilon_{t+1}, t = 1, 2, \cdots, \tag{6.1}$$

where X_t and ϵ_t are d-dimensional vectors, $\mathbb{X}_t = (X_t, \cdots, X_{t+1-p})$ is a $d \times p$ matrix and m is an unknown d-dimensional measurable function. Our objective is to approximate this unknown function by a suitable multivariate feedforward network. Therefore, we shall need to minimize a Multivariate Least Squares Problem with unknown covariance matrix Σ of ϵ_t. Roughly speaking, the problem is in general formulated as follows.
The data are approximated by

$$X_{t+1} = f(X_t, \cdots, X_{t+1-p}, \theta) + \epsilon_{t+1}, t = 1, 2, \cdots, \tag{6.2}$$

when they were actually generated by

$$X_{t+1} = m(X_t, \cdots, X_{t+1-p}) + \epsilon_{t+1}, t = 1, 2, \cdots \tag{6.3}$$

and one is supposed to solve the minimization problem

$$Q_n(\theta, \Sigma) = \sum_{t=0}^{n-1} (X_{t+1} - f(\mathbb{X}_t, \theta))' \Sigma^{-1} (X_{t+1} - f(\mathbb{X}_t, \theta)). \tag{6.4}$$

The complexity of the problem can be regarded as that of solving a Multivariate Least Squares Problem, where, in general, Σ^{-1} is unknown and not always easy to estimate. Nevertheless, if we were give a consistent estimate $\hat{\Sigma}$, we could have solved the following problem, i.e. minimize

$$Q_n(\theta, \hat{\Sigma}) = \sum_{t=0}^{n-1} (X_{t+1} - f(\mathbb{X}_t, \theta))' \hat{\Sigma}^{-1} (X_{t+1} - f(\mathbb{X}_t, \theta)), \tag{6.5}$$

with $\hat{\Sigma}^{-1}$ being a known matrix. For sake of simplicity, in the following lines, we make some (realistic) assumptions on Σ^{-1} to overcome the problem caused by its estimation.
On this way, if we assume Σ to be positive semi-definite(which should be the case for any covariance matrix), from Linear algebra, one can construct the square root $\Sigma^{-1/2}$ of Σ^{-1} with the property $\Sigma^{-1/2'} = \Sigma^{-1/2}$. By doing so the optimization problem can be reformulated in the following way:

$$\begin{aligned} Q_n(\theta, \Sigma) &= \sum_{t=0}^{n-1} (X_{t+1} - f(\mathbb{X}_t, \theta))' \Sigma^{-1/2} \Sigma^{-1/2} (X_{t+1} - f(\mathbb{X}_t, \theta)) \\ &= \sum_{t=0}^{n-1} (\Sigma^{-1/2}(X_{t+1} - f(\mathbb{X}_t, \theta)))' (\Sigma^{-1/2}(X_{t+1} - f(\mathbb{X}_t, \theta))) \\ &= \sum_{t=0}^{n-1} (\Sigma^{-1/2} X_{t+1} - \Sigma^{-1/2} f(\mathbb{X}_t, \theta)))' (\Sigma^{-1/2} X_{t+1} - \Sigma^{-1/2} f(\mathbb{X}_t, \theta))) \end{aligned}$$

From this reformulation one can observe that a form similar to the Euclidean distance appears and this is our motivation for formulating the problem as that of minimizing the least squares Euclidean norm. For sake of simplicity, we limit ourselves to the type of problem where the $\Sigma \equiv I_d$ (d-dimensional identity matrix), i.e. minimize,

$$\begin{aligned} Q_n(\theta) &= \sum_{t=0}^{n-1} (X_{t+1} - f(\mathbb{X}_t, \theta))' (X_{t+1} - f(\mathbb{X}_t, \theta)) && (6.6) \\ &= \sum_{t=0}^{n-1} \|X_{t+1} - f(\mathbb{X}_t, \theta)\|^2. && (6.7) \end{aligned}$$

We can rewrite this as

$$\begin{aligned} Q_n(\theta) &= \sum_{t=0}^{n-1} \sum_{u=1}^{d} (X_{t+1,u} - f_u(\mathbb{X}_{t,u}, \theta_u))^2 && (6.8) \\ &= \sum_{t=0}^{n-1} Q_{t,M}. && (6.9) \end{aligned}$$

By doing so we implicitly have to solve the minimization problem:

$$\min_{\theta} \mathbb{E} Q_{t,M} = \min_{\theta} Q(\theta). \tag{6.10}$$

As we have already done for the least squares and the weighted least squares in the one dimensional problem, we need to find a solution of:

$$\frac{\partial Q_n}{\partial \theta_i}(\theta) = 0 \text{ for all } i. \tag{6.11}$$

To achieve this goal, we need once more to control the asymptotic behavior of the Taylor expansion of $S_n(\theta)$ and hence the need of assumptions similar to those we have stated up to now. Let us first define

$$V_{n,S} = \nabla^2 S_n(\theta_M)$$
$$T_{n,S} = \nabla^2 S_n(\theta_M^*) - V_{n,S}$$

with θ_M being the unique global minimum of $Q(\theta)$ and θ_M^* defined as θ^* in the least squares problem relative to θ_M.
Basically we have

$$\frac{\partial Q_{t,M}}{\partial \theta_{u,i}} = -\frac{\partial f_u}{\partial \theta_{u,i}}(\mathbb{X}_{t,u}, \theta_u)(X_{t+1,u} - f_u(\mathbb{X}_{t,u}, \theta_u)) \tag{6.12}$$

and

$$\frac{\partial^2 Q_{t,M}}{\partial \theta_{u,i} \partial \theta_{h,j}} = \begin{cases} 0 & \text{if } h \neq u \\ -\frac{\partial^2 f_u}{\partial \theta_{u,i} \partial \theta_{u,j}}(\mathbb{X}_{t,u}, \theta_u)(X_{t+1,u} - f_u(\mathbb{X}_{t,u}, \theta_u)) & \\ +\frac{\partial f_u}{\partial \theta_{u,i}}(\mathbb{X}_{t,u}, \theta_u)\frac{\partial f_u}{\partial \theta_{u,j}}(\mathbb{X}_{t,u}, \theta_u) & \text{otherwise.} \end{cases}$$

6.1.1 Consistency and Asymptotic Normality

Let us now state some basic assumptions on $\{X_t\}$ and derive some consequences

A. 6.1 *Assume that $\{X_t\}$ is strictly stationary, α- mixing with mixing rate $\alpha(k) \leq ak^{-\beta}$ for some $a > 0$ and some $\beta > 1$.*

From this assumption, it follows that the marginals $\{X_{t,u}, u = 1, \cdots, d\}$ are also strictly stationary, α-mixing with mixing rate not greater than $\alpha(k)$.

A. 6.2 *1. Assume that X_t is strictly stationary α-mixing and $\sup_u \mathbb{E}|X_t^u|^{2\gamma} < \infty$ for some $\gamma > 2$.*

2. $\exists\, C_j > 0$ such that $\sup_u \mathbb{E}(|\varepsilon_t^u|^j \mid \mathbb{X}_t^u = x) \leq C_j < \infty$ for $j = 1, \cdots, 4$.

3. $m_k : \mathbb{R}^p \to \mathbb{R}$ are continuous and $\exists\ \kappa \geq 0$ such that $\sup_k |m_k(x)| \leq \inf_k \sum_{i=1}^p |x_i^k| + \kappa$.

A. 6.3 *Consider **A.**5.2 to hold for each network function f_u.*

A. 6.4 *Q has its unique global minimum at θ_M which is an interior point of Θ_M and $\nabla^2 Q(\Theta_M) = A_Q(\Theta_M)$ is positive definite.*

Proposition 6.5 *Let us consider **A.**6.1 to **A.**6.4 to hold. Then,*

1. $\lim_{n\to\infty} \frac{1}{n}\frac{\partial Q_n}{\partial \theta_{i,u}}(\theta_M) = 0$ a.s., for $i = 1, \cdots, l(u); u = 1, \cdots, d$.

2. $\lim_{n\to\infty} \frac{1}{n} V_{n,Q}(\theta_M) \to V$ *a.s., where* V *is a positive definite matrix.*

3.

$$\limsup_{n\to\infty} \sup_{\delta\to 0} \left(\frac{|T_{n,Q}(\theta_M^*)_{i,j}|}{n\delta} \right) < \infty \text{ a.s. for all } i,j = 1,\cdots,l(u), u = 1,\cdots,d.$$

Proof: The proof is similar to that of Proposition 5.4. ∎

A. 6.6 ***A.****5.8 holds for* $\{X_t\}$ *(compare Chapter 5, Section 5.1.2).*

A. 6.7 *For each u,* $f_u(x,\theta)$*, its first, and second partial derivative w.r.t.* θ*, are measurable w.r.t.* x *and uniformly continuous in a neighborhood of* θ_M *for every* x*.*

Theorem 6.8 *Let* **A.***6.1 to* **A.***6.7 hold, then*

$$n^{1/2}(\theta_n - \theta_M) \to \mathcal{N}\left(0, \Sigma(\theta_M)\right),$$

where

$$\begin{aligned}
\Sigma &= A^{-1}(\theta_M) B(\theta_M) A^{-1}(\theta_M), \\
A &= \left(\frac{\partial^2 \mathbb{E} Q_{t,M}(\theta_M)}{\partial\theta_i \partial\theta_j} \right)_{i,j} \\
B(\theta_M) &= \mathbb{E} \nabla Q_{t,M}(\theta_M) \nabla Q_{t,M}(\theta_M)'.
\end{aligned}$$

Proof: **A.**6.1 to **A.**6.7 provide analogous of Theorem 5.7 and Proposition 5.10 with the slight modification that one needs to use a Multivariate Central Limit Theorem for stationary α-mixing processes(see, e.g. Billingsley [9]) and the Ergodic Theorem applied to stationary vector-valued processes. The proof follows similarly to that of Theorem 5.13. ∎

6.2 Nonlinear Multivariate Weighted Least Squares

Let $\{X_t\}_{t\geq 0}$ be a multivariate nonlinear autoregressive process with switching

$$X_t = \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_{t,k}) \text{ with } S_{t,k} = \begin{cases} 1 & \text{for } Q_t = k \\ 0 & \text{else} \end{cases} \quad (6.13)$$

6.2.1 Preliminaries

In this context, the Weighted Least Squares is defined as follow

$$Q_{n,G} = \sum_{t=1}^{n} q_{t,G},$$

with

$$q_{t,G} = \sum_{l=1}^{L} P_t^l e_t^l,$$

$$\begin{aligned} e_t^l &= \|X_{t+1} - f(\mathbb{X}_t, \theta^l)\|^2 \\ &= \sum_{u=1}^{q} (X_{t+1,u} - f(\mathbb{X}_{t,u}, \theta_u^l))^2 \end{aligned}$$

and

$$P_t^l = \frac{\exp\left(-\beta \sum_{j=t-\tau}^{t+\tau} e_j^l\right)}{\sum_{\lambda=1}^{L} \exp\left(-\beta \sum_{j=t-\tau}^{t+\tau} e_j^\lambda\right)}.$$

Again we write

$$g(t,\lambda) = (-\beta \sum_{j=t-\tau}^{t+\tau} e_j^\lambda).$$

It follows that

$$\frac{\partial P_t^l}{\partial \theta_{i,u}^\lambda} = \begin{cases} P_t^l(1 - P_t^l)\frac{\partial g(t,l)}{\partial \theta_{i,u}^l} & \text{if } \theta_{i,u}^\lambda = \theta_{i,u}^l \\ -P_t^l P_t^\lambda \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} & \text{if } l \neq \lambda \end{cases},$$

$$\frac{\partial e_t^l}{\partial \theta_{i,u}^\lambda} = \begin{cases} -\frac{\partial f(\mathbb{X}_t, \theta_u^l)}{\partial \theta_{i,u}^l}(X_{t+1,u} - f(\mathbb{X}_{t,u}, \theta_u^l)) & \text{if } \theta_{i,u}^\lambda = \theta_{i,u}^l \\ 0 & \text{if } l \neq \lambda. \end{cases},$$

and consequently

$$\frac{\partial q_{t,G}}{\partial \theta_{i,u}^\lambda} = -\frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} P_t^\lambda q_{t,G} + \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} P_t^\lambda e_t^\lambda + P_t^\lambda \frac{\partial e_t^\lambda}{\partial \theta_{i,u}^\lambda}.$$

Hence

$$\frac{\partial^2 q_{t,G}}{\partial \theta_{i,u}^\lambda \partial \theta_{j,h}^\mu} = A_t q_{t,G} + B_t$$

with

$$A_t = \left(P_t^\lambda P_t^\mu \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} \frac{\partial g(t,\mu)}{\partial \theta_{j,h}^\mu} - \frac{\partial P_t^\lambda}{\partial \theta_{j,h}^\mu} \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} - \frac{\partial^2 g(t,\lambda)}{\partial \theta_{i,u}^\lambda \partial \theta_{j,h}^\mu} \right)$$

and

$$\begin{aligned} B_t &= -\frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} P_t^\lambda P_t^\mu \left(\frac{\partial g(t,\mu)}{\partial \theta_{j,h}^\mu} e_t^\mu + \frac{\partial e_t^\mu}{\partial \theta_{j,h}^\mu} \right) + \frac{\partial^2 g(t,\lambda)}{\partial \theta_{i,u}^\lambda \partial \theta_{j,h}^\mu} P_t^\lambda e_t^\lambda \\ &+ \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} \left(\frac{\partial P_t^\lambda}{\partial \theta_{j,h}^\mu} e_t^\lambda + P_t^\lambda \frac{\partial e_t^\lambda}{\partial \theta_{j,h}^\mu} \right) + \frac{\partial P_t^\lambda}{\partial \theta_{j,h}^\mu} \frac{\partial e_t^\lambda}{\partial \theta_{i,u}^\lambda} + P_t^\lambda \frac{\partial^2 e_t^\lambda}{\partial \theta_{i,u}^\lambda \partial \theta_{j,h}^\mu}. \end{aligned}$$

Now let us consider different situations.

1. If $\lambda = \mu$ and u= h, it follows that

$$A_t = P_t^\lambda (2P_t^\lambda - 1) \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} \frac{\partial g(t,\lambda)}{\partial \theta_{j,u}^\lambda} - \frac{\partial^2 g(t,\lambda)}{\partial \theta_{i,u}^\lambda \partial \theta_{j,u}^\lambda}$$

and

$$\begin{aligned} B_t &= \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} P_t^\lambda \left((1 - P_t^\lambda) \frac{\partial^2 e_t^\lambda}{\partial \theta_{j,u}^\lambda} - P_t^\lambda \frac{\partial g(t,\lambda)}{\partial \theta_{j,u}^\lambda} e_t^\lambda \right) \\ &+ P_t^\lambda (1 - P_t^\lambda) \frac{\partial g(t,\lambda)}{\partial \theta_{j,u}^\lambda} \left(\frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} + \frac{\partial^2 e_t^\lambda}{\partial \theta_{i,u}^\lambda} \right) \\ &+ P_t^\lambda \left(\frac{\partial^2 g(t,\lambda)}{\partial \theta_{i,u}^\lambda \partial \theta_{j,u}^\lambda} e_t^\lambda + \frac{\partial^2 e_t^\lambda}{\partial \theta_{i,u}^\lambda \partial \theta_{j,u}^\lambda} \right) \end{aligned}$$

2. If $\lambda = \mu$ and $u \neq h$, it follows that

$$A_t = P_t^\lambda (2P_t^\lambda - 1) \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} \frac{\partial g(t,\lambda)}{\partial \theta_{j,h}^\lambda}$$

and

$$\begin{aligned} B_t &= -\frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} P_t^\lambda P_t^\lambda \left(\frac{\partial g(t,\lambda)}{\partial \theta_{j,h}^\lambda} e_t^\lambda + \frac{\partial^2 e_t^\lambda}{\partial \theta_{i,u}^\lambda \partial \theta_{j,h}^\lambda} \right) + P_t^\lambda \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} \frac{\partial e_t^\lambda}{\partial \theta_{j,h}^\lambda} \\ &+ P_t^\lambda (1 - P_t^\lambda) \frac{\partial g(t,\lambda)}{\partial \theta_{j,h}^\lambda} \left(\frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^\lambda} + \frac{\partial e_t^\lambda}{\partial \theta_{i,u}^\lambda} \right). \end{aligned}$$

3. If $\lambda \neq \mu$ it follows that

$$A_t = 2P_t^{\lambda} P_t^{\mu} \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^{\lambda}} \frac{\partial g(t,\mu)}{\partial \theta_{j,h}^{\mu}}$$

and

$$\begin{aligned} B_t &= -\frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^{\lambda}} P_t^{\lambda} P_t^{\mu} \left(\frac{\partial g(t,\mu)}{\partial \theta_{j,h}^{\mu}} e_t^{\mu} + \frac{\partial e_t^{\mu}}{\partial \theta_{j,h}^{\mu}} \right) - \\ &+ \frac{\partial g(t,\lambda)}{\partial \theta_{i,u}^{\lambda}} P_t^{\lambda} P_t^{\mu} \frac{\partial g(t,\mu)}{\partial \theta_{j,h}^{\mu}} e_t^{\lambda} - \frac{\partial g(t,\mu)}{\partial \theta_{j,h}^{\mu}} P_t^{\lambda} P_t^{\mu} \frac{\partial e_t^{\lambda}}{\partial \theta_{j,h}^{\mu}}. \end{aligned}$$

Finally,

$$\begin{aligned} \frac{\partial^3 q_{t,G}}{\partial \theta_{i,u}^{\lambda} \partial \theta_{j,h}^{\mu} \theta_{k,z}^{\gamma}} &= q_{t,G} \left(\frac{\partial A_t}{\partial \theta_{k,z}^{\gamma}} - \frac{\partial g(t,\gamma)}{\partial \theta_{k,z}^{\gamma}} P_t^{\gamma} \right) \\ &+ A_t P_t^{\gamma} \left(\frac{\partial g(t,\gamma)}{\partial \theta_{k,z}^{\gamma}} e_t^{\gamma} + \frac{\partial e_t^{\gamma}}{\partial \theta_{k,z}^{\gamma}} \right) + \frac{\partial B_t}{\partial \theta_{k,z}^{\gamma}}. \end{aligned}$$

6.2.2 Consistency and Asymptotic Normality

Let us state some assumptions.

A. 6.9 ***A.*** *5.16 holds (compare Chapter 5, Section 5.2.2).*

A. 6.10 *1.* $Z_{t,k},\ k = 1, \cdots, K$ *are i.i.d. with mean* 0*,*

2. $\exists\, C_j > 0$ *such that* $\sup_{k \in I_K, h \in I_d} \mathbb{E}(|Z_{t,k,h}|^j | \mathbb{X}_t = x) \leq C_j < \infty$.

3. $m_{k,h}(\mathbb{X}), k \in I_K, h \in I_d$ *are continuous and* $\exists \kappa \geq 0, a_i \geq 0, i \in I_p$ *such that* $\sup_{k \in I_K, h \in I_d} |m_{k,h}(x)| \leq \inf_{k \in I_K, h \in I_d} \sum_{i=1}^{p} a_i |x_{i,k,h}| + \kappa$.

A. 6.11 *1.* $X_{t,k}, k = 1, \cdots, K$ *are independent multivariate NLAR(*p*)-processes.*

2. $X_{t,k}$ *is strictly stationary and* $\alpha-$*mixing for each* $k \in I_K$.

3. $\sup_{k \in I_K, h \in I_d} \mathbb{E}|X_{t,k,h}|^{4\gamma+1} < \infty$ *for some* $\gamma > 2$.

A. 6.12 ***A.*** *5.2 (compare Chapter 5, Section 5.1.2) holds for each network function* $f_{k,h}, k \in I_K, h \in I_d$.

A. 6.13 *S has its unique global minimum at θ_G which is an interior point of Θ_G and $\nabla^2 Q(\theta_g) = A_{Q_G}(\Theta_G)$ is positive definite.*

Let us define

$$V_{n,Q_G} = \nabla^2 Q_{n,G}(\theta_G)$$
$$T_{n,Q_G} = \nabla^2 Q_{n,G}(\theta_G^*) - V_{n,Q_G},$$

with θ_G being the unique global minimum of $Q_G(\theta)$ and θ_G^* defined as θ^* in the least squares problem relative to θ_G.
The two results are the analogous of proposition and theorem established earlier and the proofs follow their proofs, assuming that all intermediate states are obviously proved given our set of assumptions. Therefore, we will skip their proofs here.

Proposition 6.14 *Given **A.**6.9 to **A.**6.11 hold, it follows that*

1. $\lim_{n\to\infty} \frac{1}{n} \frac{\partial Q_{n,G}}{\partial \theta_{i,u}}(\theta_G) = 0$ *a.s., for* $i = 1, \cdots, l(u); u = 1, \cdots, d$,
2. $\lim_{n\to\infty} \frac{1}{n} V_{n,Q_G}(\theta_G) \to V$ *a.s. where V is a positive definite matrix and*
3. $$\lim_{n\to\infty} \sup_{\delta\to 0} \left(\frac{|T_{n,Q_G}(\theta_G^*)_{i,j}|}{n\delta} \right) < \infty \text{ a.s. for all } i, j = 1, \cdots, l(u), u = 1, \cdots, d.$$

The Asymptotic normality of the parameter estimate is formulated as follows,

Theorem 6.15 *Let **A.**6.9 to **A.**6.13 hold, then*

$$n^{1/2}(\theta_n - \theta_G) \to \mathcal{N}\left(0, \Sigma(\theta_G)\right),$$

where

$$\Sigma = A^{-1}(\theta_G) B(\theta_G) A^{-1}(\theta_G),$$
$$A = \left(\frac{\partial^2 \mathbb{E} q_{t,G}(\theta_G)}{\partial \theta_i \partial \theta_j} \right)_{i,j}$$
$$B(\theta_G) = \mathbb{E} \nabla q_{t,G}(\theta_G) \nabla q_{t,G}(\theta_G)'.$$

Beside all these theoretical considerations, the question we need to answer is how to solve the problem numerically? For this aim, we have to relax our assumption and use the result by White [91]. In fact, under some special conditions, he proves the convergence and asymptotic normality of Backpropagation for this type of network function.
As for the Least Squares Problem, we have spoken about the existence of a consistent sequence of estimates for the Weighted Least Squares Problem. But we have never said how we can compute these estimates. In the coming section, based on the work done by White, we will relax some of our assumptions and prove that Back-propagation can help to solve this problem from a numerical point of view.

7 A Numerical Procedure: Backpropagation

Here, we are concerned with the general problem of finding algorithm that allows to retrieve the parameters of a model given the cost function that we have to optimize. For this purpose, we will present a version of the stochastic approximation algorithm that helps us to numerically solve the problem of weighted least square as presented in the previous chapter. Basically the stochastic approximation theory goes back to Robbins and Monro, but here we shall follow the approach by White [88], who used the results by Ljung to derive convergence of gradient descent in Neural Networks context.
In this chapter we strengthen our model assumptions, and therefore, consider a regression model instead of autoregression model, i.e.

$$Y_t = m(\mathbb{X}_t) + \sigma(\mathbb{X}_t)\epsilon_t \tag{7.1}$$

with i.i.d. $\mathbb{X}_t, \epsilon_t$ instead of the model defined in equation 5.1. Then, the $Z_t = (Y_t, \mathbb{X}_t')'$ are i.i.d. random vectors. We want to simplify our considerations, as our main goal is to show that Backpropagation also work for mixture models. We are sure that the results can be extended to the α-mixing time series case 5.1 as for the case $K = L = 1$ for which White [88] has already shown the convergence of Backpropagation. Moreover, to avoid lengthy technical expositions, we restrict ourselves in this chapter to bounded i.i.d. random variables.
Let us then recall the conditions in [88] and derive the result for the Weighted Least Squares Problem in the regression situation.

7.1 Convergence of Backpropagation

Before we present the convergence theorem, let us state a proposition which can be considered as the base of the proof of the convergence of the numerical procedure. Let us first recall some useful definitions.

Definition 7.1.1 *For $\Phi \subseteq \mathbb{R}^d$ and θ_n a sequence of vectors, $\theta_n \to \Phi$ means that*

$$\inf_{\theta \in \Phi} |\theta - \theta_n| \to 0 \text{ as } n \to \infty. \tag{7.2}$$

One will write $\theta_n \to \infty$ if $|\theta_n| = \infty$.
Finally, for $\theta^* \in \mathbb{R}^d$, and $\epsilon > 0$, $S_\epsilon^* \equiv \{\theta : |\theta - \theta^*| < \epsilon\}$.

All over this chapter we consider $V = (V_1, \cdots, V_d)'$

Proposition 7.1 *(White, 1989)*
Let Z_n be a sequence of i.i.d random $v \times 1$ vectors such that $|Z_n| \leq \Delta < \infty$. Let

$m : \mathbb{R}^v \times \mathbb{R}^s \longrightarrow \mathbb{R}^s$ be continuously differentiable on $\mathbb{R}^v \times \mathbb{R}^s$ and suppose that for each $\theta \in \mathbb{R}^s$

$$M(\theta) \equiv \mathbb{E}(m(Z_n, \theta))$$

exists. Let $a_n \in \mathbb{R}^+$ be a decreasing sequence such that

$$\sum_{n=1}^{\infty} a_n = \infty, \lim_{n\to\infty} \sup(a_n^{-1} - a_{n-1}^{-1}) < \infty, \text{ and } \sum_{n=1}^{\infty} a_n^d < \infty \text{ for some } d > 1.$$

Define the recursive procedure

$$\tilde{\theta}_n = \tilde{\theta}_{n-1} + a_n m(Z_n, \tilde{\theta}_{n-1})(n = 1, 2, \cdots)$$

where $\tilde{\theta}_0 \in \mathbb{R}^s$ is arbitrary.

1. *Suppose that there exists $Q : \mathbb{R}^s \longrightarrow \mathbb{R}$ twice continuously differentiable such that $\nabla Q(\theta)' M(\theta) \leq 0$ for all $\theta \in \mathbb{R}^s$. Then either $\tilde{\theta}_n \longrightarrow \Theta^* \equiv \{\theta : \nabla Q(\theta)' M(\theta) = 0\}$ or $\tilde{\theta}_n \longrightarrow \infty$ with probability one.*

2. *Suppose that $\theta^* \in \mathbb{R}^s$ is such that $\mathbb{P}\left[\tilde{\theta}_n \longrightarrow S_\varepsilon^*\right] > 0$ for all $\varepsilon > 0$. Then $M(\theta^*) = 0$. If, in addition, M is continuously differentiable in a neighborhood of θ^* with $\nabla M^* \equiv \nabla M(\theta^*)$ finite and*

   $$J^* = \mathbb{E}(m(Z_n, \theta^*) m(Z_n, \theta^*)')$$

   is finite and positive definite, then the eigenvalues of ∇M^* all lie in the left half-plane.

3. *Suppose that the conditions of (1.) hold, that $M(\theta) = -\nabla Q(\theta)$, that $Q(\theta)$ has isolated stationary points (strong local minimum) and that the conditions of part (2.) hold for each $\theta^* \in \Theta^* = \{\theta : \nabla Q(\theta) = 0\}$. Then as $n \longrightarrow \infty$, either $\tilde{\theta}_n$ tends to a local minimum of $Q(\theta)$ with probability 1 or $\tilde{\theta}_n \longrightarrow \infty$ with probability 1.*

Proof: This proposition is due to White [88], but essentially based on the work done by Ljung [60]. Consequently, one should refer to the latter for a better understanding of the proposition. ∎

Having this proposition, we are now ready to present the asymptotic properties of the Backpropagation under our considerations. First, let us present the model we are concerned with. Let

$$\begin{aligned} Q_G(\theta^1, \cdots, \theta^l) &= \mathbb{E}(\sum_{l=1}^{L} P_n^l e_n^l) && (7.3) \\ &= \mathbb{E} g(Z_n, \theta^1, \cdots, \theta^l). && (7.4) \end{aligned}$$

with

$$e_t^l = (\ Y_t - f_l(\ \mathbb{X}_t, \theta^l))^2$$

and

$$P_t^l = \frac{\exp\left(\ -\beta \sum_{j=t-\tau}^{t+\tau} e_n^l\right)}{\sum_{\lambda=1}^{L} \exp\left(\ -\beta \sum_{j=t-\tau}^{t+\tau} e_j^\lambda\right)}$$

as defined in Chapter 5. Keeping this equation in mind, let us give a set of sufficient conditions which may allows us to derive the consistency of Backpropagation.

A. 7.2 *1. We have a training sequence $\{Z_t\}$ of i.i.d. random vectors satisfying $|Z_t| \leq \triangle < \infty a.s.$*

2. For each network n_l, we have the following:

$$f(X; \theta^l) = F_l(\nu_0^l + \sum_{h=1}^{H(l)} \nu_h^l \psi_h(\omega_{h,0}^l + \sum_{i=1}^{p} X_i \omega_{h,i}^l)),$$

where F_l is defined on $\mathbb{R}$ and the activation functions

$$\psi_h : \mathbb{R} \to \mathbb{I} \equiv [0, 1]$$

are continuously differentiable of order 2 on $\mathbb{R}$; $\theta^l = (\nu^l, \omega^l) \in \mathbb{R}^{s(l)}$, $\quad s(l) = H(l)(2 + p) + 1, H(l)$ is the number of hidden neurons in the network n_l for $l = 1, \cdots, L$.

3. $a_n \in \mathbb{R}^+$ is a decreasing sequence such that:

(a) $\sum_{n=1}^{\infty} a_n = \infty$

(b) $\lim_{n\to\infty} \sup(a_n^{-1} - a_{n-1}^{-1}) < \infty$

(c) $\sum_{n=1}^{\infty} a_n^d < \infty$ *for some* $d > 1$

Let make some comments on **A.** 7.2.
Although **A.**7.2 -1. is somehow restrictive from the theoretical point of view(since it excludes the common Gaussian process), this assumption is reasonable for all practical purposes because it does not make sense to work with processes that take values at infinity.

In the current work we choose the activation functions to be identical and define them to be a type of sigmoid, for example the logistic function

$$\phi(u) = \frac{1}{1+e^{-u}}. \tag{7.5}$$

Conditions **A.**7.2-3a and **A.**7.2-3c are generalized versions of the conditions which Robbins and Monro used in their pioneering work on stochastic approximation theory. Remark that **A.**7.2-3a is a necessary condition for the convergence of $\tilde{\theta}_n$ even if there is no random error. Beside this, we can observe that a_n should not to be too large, otherwise the random error will prevent the convergence. It turns out that **A.**7.2-3c asymptotically damps the effect of experimental errors. Assumption **A.**7.2-3b was first used by Ljung [60] for technical reason in the proof of his results. In general assumption **A.**7.2 holds for a very wide choice of a_n, the latter is usually considered as $a_n = an^{-\alpha}(0 < \alpha \leq 1)$. The leading choice is with $\alpha = 1$.

Lemma 7.3 *Under the **A.** 7.2-1 and **A.** 7.2-2,*

$$\begin{aligned} Q_G(\theta^1, \cdots, \theta^l) &= \mathbb{E}(\sum_{l=1}^{L} P_n^l e_n^l) && (7.6)\\ &= \mathbb{E}g(Z_n, \theta^1, \cdots, \theta^l) && (7.7) \end{aligned}$$

is $\mathcal{C}^2$ on $\mathbb{R}^u$ w.r.t. θ, moreover

$$D^k Q_G(\theta^1, \cdots, \theta^l) = \mathbb{E}D^k(\sum_{l=1}^{L} P_n^l e_n^l) \ \textit{for } k = 1, 2. \tag{7.8}$$

where D^k are the k'th partial differentiations w.r.t. the parameters.

Proof: In this proof we essentially need to apply a theorem on the derivation of parametric integrals, for example the theorem in Schmets [81] page 46.
From **A.** 7.2-2, it follows that $\sum_{l=1}^{L} P_n^l e_n^l = g(z, \theta)$ is twice continuously differentiable on $\mathbb{R}^u \times \mathbb{R}^s$, from which we can conclude that $g(z, \theta)$ is measurable with respect to z and twice continuously differentiable w.r.t θ on $\mathbb{R}^s$ for each z.
If we now consider a connected component of $\mathbb{R}^s$, i.e. $\mathbb{R}^s$ itself and choose $\theta_c = 0$ in this connected component, then we obtain (by **A.** 7.2-1) that $|g(z, \theta_c)|$ is dominated by an integrable function w.r.t to the probability measure. Similarly, if we also take $\theta_c \neq 0$(with the help of **A.** 7.2-1), it is easy to show that $\|\nabla g(z, \theta_c)\|$ is dominated by an integrable function.
Let us now consider a compact $K \subset \mathbb{R}^s$. For $\theta \in K$, continuity of $\nabla^2 g(z, \theta)$ on the compact subset(from **A.** 7.2.1) of $\mathbb{R}^u$ containing z implies that $\nabla^2 g(z, \theta)$ is bounded. Hence, $\|\nabla^2 g(z, \theta)\|$ is dominated by an integrable function and the proof of the lemma follows. ∎

Theorem 7.4 *Let us assume the conditions of* **A.** *7.2 are satisfied, define the Backpropagation estimator*

$$\tilde{\theta}_n = \tilde{\theta}_{n-1} - a_n \nabla g(\tilde{\theta}_{n-1}), \quad n = 1, 2, \cdots, \tag{7.9}$$

where θ_0 is arbitrary. Then either

$$\tilde{\theta}_n \to \Theta^* = \{\theta : \mathbb{E}(\nabla g(\theta) = 0\} \text{ w.p.1. or } \tilde{\theta}_n \to \infty \text{ w.p.1.}.$$

If in addition, $Q_G(\theta)$ has isolated stationary points such that

$$J^* = \mathbb{E}(\nabla g(Z_n, \theta^*)' \nabla g(Z_n, \theta^*))$$

is positive definite for each $\theta^* \in \Theta^*$, then either $\tilde{\theta}_n$ converges to a local minimum of $Q_G(\theta)$ w.p.1 or $\tilde{\theta}_n \to \infty$.

Proof: For this proof we apply Lemma 7.3 and Proposition 7.1. Now Consider $m(z, \theta) = \nabla g(z, \theta)$,with $g(z, \theta)$ as defined in the proof of the previous lemma. **A.** 7.2.1 ensures that $\{Z_n\}$ are i.i.d. uniformly bounded. From **A.** 7.2.2, $\nabla g(z, \theta)$ is continuously differentiable and the lemma provides that for each $\theta \in \mathbb{R}^s$, $M(\theta) = \mathbb{E}\nabla g(z, \theta)$ exists. The conditions on a_n are provided by 7.2.3.
If we take

$$Q(\theta) = -\mathbb{E} \sum_{l=1}^{L} P_n^l e_n^l = -\mathbb{E} g(z, \theta),$$

by Lemma 7.3, $Q(\theta)$ is twice continuously differentiable and

$$\theta \in \mathbb{R}^s, \nabla Q(\theta)' M(\theta) \leq 0$$

for all $\theta \in \mathbb{R}^s$. Then the conclusion of Proposition 7.1.2 holds and therefore the first part of the theorem.
By Lemma 7.3, $\nabla M(\theta^*)$ is finite and by assumption J^* is positive definite. Hence, the conditions of Proposition 7.1-2 hold. By Proposition 7.1-3, $\tilde{\theta}_n$ converges to a local minimum of $Q(\theta)$. ∎

7.1.1 Asymptotic Normality

In this section we essentially recall the assumptions by White [88], assumptions that he uses to establish the asymptotic normality of the Backpropagation.

Proposition 7.5 *Let the conditions of Proposition 7.1.(1, 2) hold and suppose that $\|g(z, \theta)\| < \Delta < \infty$ a.s. for all $\theta \in \mathbb{R}^s$. Let λ^* be the maximum value of the real part of the eigenvalues of $\nabla M(\theta^*)$ and suppose that $\lambda^* < \frac{-1}{2}$. Define $J(\theta) = var[g(z, \theta)]$ and suppose that J is continuous in a neighborhood of θ^*. Set $J^* =$*

$J(\theta^*)$ and $a_n = \frac{1}{n}$. Then the sequence of the random element $T_n(a)$ of $\mathcal{C}_{\mathbb{R}^l}[0,1]$, with sup norm defined by

$$T_n(a) = n^{-1/2} S_{[na]} + (na - [na]) n^{-1/2} (S_{[na]+1} - S_{[na]}), a \in [0,1]$$

(where $S_n = (\tilde{\theta}_n - \theta^*)$), converges to a Gaussian Markov process G with

$$G(a) = \exp[(\ln(a)(I + \nabla M(\theta^*)] \times \int_{(0,a]} \exp[-(\ln(s)(I + \nabla M(\theta^*)] dW(s), a \in (0,1],$$

where W is a Brownian motion in $\mathbb{R}^l$ with

$$W(0) = 0, \mathbb{E}W(1) = 0 \quad \textit{and} \quad \mathbb{E}W(1)W(1)' = J(\theta^*).$$

In particular,

$$n^{1/2}(\tilde{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, F^*),$$

where

$$F^* = \int_{(0,1]} \exp[-(\ln(s)(I + \nabla M(\theta^*)] J^* \exp[-(\ln(s)(\nabla M(\theta^*)' + I)] ds$$

is the unique solution of the equation

$$(\nabla M(\theta^*) + I/2) F^* + F^* (\nabla M(\theta^*)' + I/2) = -J^*.$$

When $\nabla M(\theta^*)$ is symmetric, $F^* = PHP^{-1}$, where P is the orthogonal matrix such that $P\Omega P^{-1} = -\nabla M(\theta^*)$ with Ω the diagonal matrix containing the (real) eigenvalues $(\lambda_1, \cdots, \lambda_l)$ of $-\nabla M(\theta^*)$ in the decreasing order and H the $l \times l$ matrix with elements

$$H_{ij} = (\lambda_i + \lambda_j - 1) K^*_{i,j} (i, j = 1, \cdots, l),$$

where $K^* = P^{-1} J^* P$.

A. 7.6 *Let **A.** 7.2.2 hold and let F, its derivatives and ψ be bounded.*

Roughly speaking, this rules out the case(our case) in which F has to be regarded as the identity. However, taking

$$F(\lambda) = \begin{cases} \lambda \text{ for } |\lambda| \leq \bar{\Delta} \\ \text{smooth and bounded for } |\lambda| > \bar{\Delta} \end{cases} \tag{7.10}$$

allows the approximation of the identity map. Additionally, since we assume X_t to be bounded, the last approximation is always possible considering $\bar{\Delta}$ large enough, e.g. $\bar{\Delta} = CM_0$ where $C > 1$ and M_0 the maximum value of the available data.

A. 7.7 *For $n = 1, 2, \cdots$, $a_n = \delta n^{-1}, \delta > 0$*

Theorem 7.8 *Let* **A.** *7.2.1,* **A.***7.6,* **A.***7.7 be given and define $\tilde{\theta}_n$ as in equation (7.9). Suppose that $\tilde{\theta}_n \to \theta^* a.s.$, θ^* an isolated stationary point of $Q(\theta)$ with J^* positive definite. Further, suppose that $a > (2\lambda^*)^{-1}$, where λ^* is the smallest eigenvalue of $\nabla^2 Q^* \equiv \nabla^2(\theta^*)$. Then, $T_n(a)$ as defined in the previous proposition converges in distribution to a Gaussian Markov process G with*

$$G(a) = \delta \exp((\ln a)[I - \delta\nabla^2 Q_G(\theta^a st)] \int_{(0,a]} \exp[(\ln s)(\delta\nabla^2 Q_G(\theta^*) - I)]dW(s),$$

with $W(0) = 0, \mathbb{E}(W(1)) = 0$, and $\mathbb{E}(W(1)W(1)') = J(\theta^*)$.
In particular

$$n^{1/2}(\tilde{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, F^*) \tag{7.11}$$

with $F^* = PHP^{-1}$, where P is the orthogonal matrix such that $P\Omega P^{-1} = \nabla^2 Q^*$, with Ω the diagonal matrix containing the eigenvalues $(\lambda_1, \cdots, \lambda_l)$ of $\nabla^2 Q^*$ in decreasing order and H the $s \times s$ matrix with the elements

$$H_{ij} = a^2(a\lambda_i + a\lambda_j - 1)^{-1} K^*_{ij} \; i, j, = 1, \cdots, s,$$

where $K^* = P^{-1} J^* P$.

Proof: The proof is essentially based on Proposition 7.5.
Assumptions **A.**7.2-1, **A.**7.6 and **A.**7.7 suffice for the assumptions **A.**7.2-1to **A.**7.2-3 to hold. Thus, for the assumptions of Proposition 7.1.1. Since θ^* is assumed to be isolated stationary point, the first condition of Proposition 7.1.2 holds. The remaining conditions hold by a suitable use of Lemma 7.3 and given that $J(\theta^*)$ is assumed positive definite. Assumptions **A.**7.2.1 and **A.**7.6 ensure that $\|\nabla g(z, \theta)\|$ is a.s. uniformly bounded for each $\theta \in \mathbb{R}^s$ and the conditions on λ^* are given by assumption. Finally the continuity of J in a neighborhood of θ^* is provided by Lemma 7.3. ∎

The results presented in this section are proved under the main assumption of bounded i.i.d. random variables, assumptions that do not fit well to our case since nonlinear Autoregressive processes are dependent by nature.
To make use of the dependence structure of the data and overcome the boundedness assumption, we can refer to the work done by Métivier and Priouret [69] where they make a Markov assumption on the observed process and required strong assumption like, e.g., the existence of all moments for the Markov process and some regularity assumptions on the gradient of the cost function. Under their assumptions they proved the convergence of the Backpropagation algorithm as defined in equation7.9.

8 Excursion to Tests in Changepoints Detection

In this Chapter we recall the importance of tests for changepoint problem. We then state the general problem for our model and derive a first test for validating change in the dynamic of a parametric Nonlinear Autoregressive model. Here we consider that the autoregressive and volatility functions are correctly specified by suitable Feedfoward Network with finite amount of hidden units, i.e. we are no longer in a nonparametric setting.

8.1 Generalities

Under the null hypothesis, that assuming there is no change in the dynamic of the observed process, the model we defined in Chapter 2, equation 2.2, i.e.

$$X_t = \sum_{k=1}^{K} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\epsilon_{t,k}) \text{ with } S_{t,k} = \begin{cases} 1 & \text{for } k = Q_t \\ 0 & \text{otherwise,} \end{cases} \tag{8.1}$$

can be written as

$$X_t = m(\alpha, \mathbb{X}_{t-1}) + \sigma(\beta, \mathbb{X}_{t-1})\varepsilon_t \tag{8.2}$$

with ε_t i.i.d. random variables and $\mathbb{X}_{t-1} = (X_{t-1}, \cdots, X_{t-p})$. Under some regularity assumption on the autoregressive function m, the volatility function σ and some mixing condition on $\{ X_t \}$ one can prove the asymptotic normality of the parameter estimate. Instead of facing the not easy to solve problem of estimating the time of change by considering the autoregressive and volatility functions as well, we suggest to follow a stepwise approach. First we recommend to consider an autoregressive model without volatility function, in a next step one can consider a nonlinear ARCH model and finally one may focus on the more complex NAR-ARCH model.
In the next section we will focus only on the first step and propose a test for estimating the change in the model assuming there is only one change.

8.2 Test for Changes in Nonlinear Autoregressive Model

Let us consider under the null hypothesis the following model

$$X_t = f(\mathbb{X}_{t-1}, \theta) + \varepsilon_t \tag{8.3}$$

where f is a given neural network.
Now define the nonlinear least square(NLLS)

$$Q_n(\theta) = \sum_t ((X_t - f(\theta, \mathbb{X}_{t-1}))^2 = \min!$$

Let consider θ_n to be the least square estimate of θ_0 the optimal parameter under the null hypothesis of no change. By solving the Least Square Equations we obtain

$$\nabla Q_n(\theta_n) = 0. \tag{8.4}$$

Providing the assumptions **A.**5.1 to **A.**5.9 of Chapter 5 one had proved that

$$\sqrt{n}(\theta_n - \theta_0) \longrightarrow \mathcal{N}(0, \Sigma(\theta)) \tag{8.5}$$

Under the alternative we can rewrite our model with changes as following

$$X_t = f(\theta_t, \mathbb{X}_{t-1}) + \varepsilon_t. \tag{8.6}$$

In general we shall make the following assumption on the residuals

A. 8.1 *The ε_t are i.i.d. random variables with zero mean, nonzero variance σ^2 and finite moments $\mathbb{E}\,|\varepsilon_t|^{2+\nu}$ for some $\nu > 0$.*

For our purpose we will assume that under the alternative the time interval can be split in two subinterval and for each interval θ_t has constant value. In other word we assume the existence of random integer $1 \leq k < n$ such that for this k,

$$\theta_1 = \theta_2 = \cdots = \theta_k \quad \text{and } \theta_{k+1} = \cdots = \theta_n \neq \theta_k.$$

Following an idea by Huskova [52], we will use the M-estimator approach to derive the test for detecting changepoint in this type of situation. In this light we reformulate the hypothesis, i.e. there is no change in the dynamic of the process as it follows

$$H_0 : \theta_{n,1} = \theta_{n,2} = \cdots = \theta_{n,n} \tag{8.7}$$

and the alternative, that is there exists exactly one change in the dynamic of the process as

$$\begin{aligned} H_1 :\ & \exists\, k, \quad 1 \leq k < n \text{ such that} \\ & \theta_{n,1} = \theta_{n,2} = \cdots = \theta_{n,k} \text{ and } \theta_{n,k+1} = \theta_{n,k+2} = \cdots = \theta_{n,n} \neq \theta_{n,k}. \end{aligned}$$

Now, let us define

$$\hat{\varepsilon}_t = X_t - f(\theta_n, \mathbb{X}_{t-1})$$

it follows by equation 8.4 that

$$\sum_{1 \leq t \leq n} \hat{\varepsilon}_t = 0 \tag{8.8}$$

that is

$$\sum_{1 \leq t \leq n} (X_t - f(\theta_n, \mathbb{X}_{t-1})) = 0 \tag{8.9}$$

Let consider $\hat{\sigma}_n^2$ a consistent estimate of the variance of ε_t, and define for each $k = 1, \cdots, n-1$ the partial sum

$$\hat{S}_{n,k} = \sum_{1 \leq t \leq k} \hat{\varepsilon}_t, \tag{8.10}$$

and in turn

$$T_n = n^{-1/2} \max_{1 \leq k < n} \left| \hat{S}_{n,k} \right|. \tag{8.11}$$

Analogously, one can define

$$S_{n,k} = \sum_{1 \leq t \leq k} \varepsilon_t, \tag{8.12}$$

Our goal is to prove that under the hypothesis

$$\frac{T_n}{\hat{\sigma}_n} \longrightarrow \sup_{0<s<1} |B(s)|$$

where $B(s)$ is a Brownian Bridge. Before we move forward in the definition of the test, we first recall the definition of the Brownian Bridge.

Definition 8.2.1 *A stochastic process $B(t)$ $0 \leq t \leq 1$ is called Brownian Bridge if $B(t)$ is a Gaussian process with $B(0) = B(1) = 0$,*

$$\mathbb{E}B(t) = 0 \tag{8.13}$$

and

$$cov(B(t), B(s)) = s(1-t) \quad 0 \leq s \leq t \leq 1. \tag{8.14}$$

Back to our goal, instead of doing the proof for $\frac{T_n}{\hat{\sigma}}$ we define

$$T_n^* = n^{-1/2} \sup_{0<s<1} \hat{S}_{n,[ns]}$$

where $[ns]$ is the integer part of ns and carry out the proof for $\frac{T_n^*}{\hat{\sigma}_n}$, for which we can now sketch the proof. By definition we have that

$$\begin{aligned}
\hat{S}_{n,[ns]} &= \hat{S}_{n,[ns]} - s\hat{S}_{n,n} \quad (\text{ by equation}8.4) && (8.15)\\
&= \hat{S}_{n,[ns]} - s\hat{S}_{n,n} - (S_{n,[ns]} - sS_{n,n}) + (S_{n,[ns]} - sS_{n,n}) && (8.16)\\
&= (\hat{S}_{n,[ns]} - S_{n,[ns]}) + s(S_{n,n} - \hat{S}_{n,n}) + (S_{n,[ns]} - sS_{n,n}) && (8.17)\\
&= T_{1,n} + T_{2,n} + T_{3,n} && (8.18)
\end{aligned}$$

where

$$T_{1,n} = \hat{S}_{n,[ns]} - S_{n,[ns]} \tag{8.19}$$

$$T_{2,n} = s(S_{n,n} - \hat{S}_{n,n}) \tag{8.20}$$

$$T_{3,n} \;=\; (S_{n,[ns]} - sS_{n,n}) \tag{8.21}$$

By the weak convergence of the partial sum as used in Theorem $2.8.4$ in [13], one has

$$\frac{T_{3,n}}{n^{1/2}\sigma} \;\longrightarrow\; B(s) \tag{8.22}$$

where $\{B(s),\ 0 < s < 1\}$ is a Brownian Bridge. For the remainder, we will apply a Taylor expansion on f and use the limit distribution of the parameter estimate to conclude.
Let us first apply a Taylor expansion on f around θ_0. we then obtain

$$f(\theta_n, \mathbb{X}_{t-1}) = f(\theta_0, \mathbb{X}_{t-1}) + \nabla f(\theta_0, \mathbb{X}_{t-1})(\theta_n - \theta_0) + o_p\left(\|\theta_n - \theta_0\|\right).$$

Taking this into, that is by considering the asymptotic and neglecting the terms of smaller order we obtain

$$\begin{aligned}
& n^{-1/2} \sum_{1\le t\le [ns]} (f(\theta_n, \mathbb{X}_{t-1}) - f(\theta_0, \mathbb{X}_{t-1})) && (8.23)\\
=\;& n^{-1/2} \sum_{1\le t\le [ns]} \nabla f(\theta_0, \mathbb{X}_{t-1})(\theta_n - \theta_0) && (8.24)\\
\cong\;& \left(n^{-1} \sum_{1\le t\le ns} \nabla f(\theta_0, \mathbb{X}_{t-1})\right) n^{1/2}(\theta_n - \theta_0) && (8.25)\\
=\;& s\left(\frac{1}{ns} \sum_{1\le t\le ns} \nabla f(\theta_0, \mathbb{X}_{t-1})\right) n^{1/2}(\theta_n - \theta_0) && (8.26)\\
\propto\;& s\left(\frac{1}{n} \sum_{1\le t\le n} \nabla f(\theta_0, \mathbb{X}_{t-1})\right) n^{1/2}(\theta_n - \theta_0). && (8.27)
\end{aligned}$$

Hence, up to a constant factor $\frac{T_{1,n}}{n^{1/2}}$ and $\frac{T_{2,n}}{n^{1/2}}$ are asymptotically equivalent. Therefore, if we prove that $\sup_{0<s<1}\left|\frac{T_{2,n}}{n^{1/2}\sigma}\right|$ converges in probability to zero we can conclude our claim. Indeed that is the case, we just need to notice by means of the asymptotic normality of θ_n and convergence of

$$\frac{1}{n} \sum_{1\le t\le n} \nabla f(\theta_0, \mathbb{X}_{t-1}),$$

imply that

$$\left|\frac{T_{n,2}}{n^{1/2}\sigma}\right| = o_p(1) \tag{8.28}$$

and the claim follows.
With all the above development we can now state the main theorem of this section, derive some of its consequences and propose a way to apply it.

Theorem 8.2 *Under the assumptions **A**.5.1 to **A**.5.9 of Chapter 5, **A**.8.1 and the null hypothesis it follows that*

$$\frac{1}{\sqrt{n}\hat{\sigma}_n} \max_{1\leq k<n} \left|\hat{S}_{n,k}\right| \longrightarrow \sup_{0<s<1} |B(s)| \tag{8.29}$$

where $\{B(s),\ 0 < s < t\}$ *is a Brownian Bridge and*

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{1\leq t\leq n} (X_t - f(\theta_n, \mathbb{X}_{t-1}))^2$$

To be complete with the proof, let us remark that we need to make use of the development that preceed the theorem. Additionally, we need to observe that $\hat{\sigma}_n^2$ is a consistent estimate of σ_n^2 and apply Slutsky's Lemma to derive the result in equation 8.29 from equation 8.22.
Practically, this theorem can be used in the following way, reject the null hypothesis if

$$\frac{1}{\sqrt{n}\hat{\sigma}_n} \max_{1\leq k<n} \left|\hat{S}_{n,k}\right| > U_{1-\alpha} \tag{8.30}$$

where $U_{1-\alpha}$ is the $(1-\alpha)$-quantile of $\sup_{0<s<1} |B(s)|$.
In this chapter we develop a test assuming only one change in the dynamic of the process and as byproduct we are also able to derive an estimate time of change. For this time of change it will be of great interest to investigate its asymptotic property. Moreover, it will be interesting to extend the results of the current chapter to the models with volatility component, that is nonlinear AR-ARCH model. Since it is limiting to assume only one change in dynamic of the process, it also important to extend the test to models with multiple changes.

9 Case Studies

In this chapter we want to compare the theory of the Hidden Markov Driven models presented in Chapter 4 and the theory of the weighted least squares developed in Chapter 5 to the reality. In this light, we first apply both approaches to computer generated data, for which we know the exact hidden structure. Later we apply the hidden Markov Driven models approach to the forecasting of daily DAX values and those of one of its main components, namely the BASF stock values.

9.1 Computer Generated Data

It is judicious for complex model classes to explore the behavior of the models in order to build up some intuition about well the models perform under well-known data structure. Since GMAR-ARCH models contain an unsupervised part in learning, in this section, we investigate whether the dynamics delivered by the models actually correspond to the true ones. In other words we investigate whether the hidden states are properly detected. For this purpose, we first consider a mixture of stationary first order autoregressive processes AR(1) to which we apply the weighted least squares techniques that were presented in Chapter 5 and alternatively we generate data from a mixture of two nonlinear AR-ARCH(1) with hidden Markov state process and apply the hidden Markov technique presented in Chapter 4 to these observations. For both cases the random residuals that we used were generated with MATLAB 6.5 and the numerical computations were achieved by some self-implemented MATLAB routines.

9.1.1 Mixture of Stationary AR(1) and Weighted Least Squares Techniques

The data used in this section consist of two different stationary first order autoregressive processes $Y_{t,1}, Y_{t,2}$ and the hidden process $S_t = (S_{t,1}, S_{t,2})$ that determines the states of the process. Having all these processes, we build up the following simple mixture model

$$X_t = S_{t,1}Y_{t,1} + S_{t,2}Y_{t,2}, \tag{9.1}$$

with

$$S_{t,1} = \begin{cases} 1 & \text{if} \quad t = 2kL+1, \cdots, (2k+1)L, \quad k \in \mathbb{N} \\ 0 & \text{otherwise,} \end{cases} \tag{9.2}$$

$$S_{t,2} = \begin{cases} 1 & \text{if} \quad t = (2k+1)L+1, \cdots, (2k+2)L, \quad k \in \mathbb{N} \\ 0 & \text{otherwise,} \end{cases} \tag{9.3}$$

where L is a given integer,

$$Y_{t,1} = \alpha_1 X_{t-1} + \epsilon_t$$

and

$$Y_{t,2} = \alpha_2 X_{t-1} + \zeta_t$$

for which ϵ_t, ζ_t are independent sequences of i.i.d. $\mathcal{N}(0,1)$ random variables.
We generate a sequence of 1200 realizations of the hidden process. This sequence determines which of the two processes is used in each time interval to generate the realization of the observed process. From the generated observations (sequence), we used the first 500 as the training set and the remainder as the validation set. For this simulation we consider the estimated instants of change and compare them with the real time of change that we have considered for generating our process. We then present the results of the simulation in the following picture.

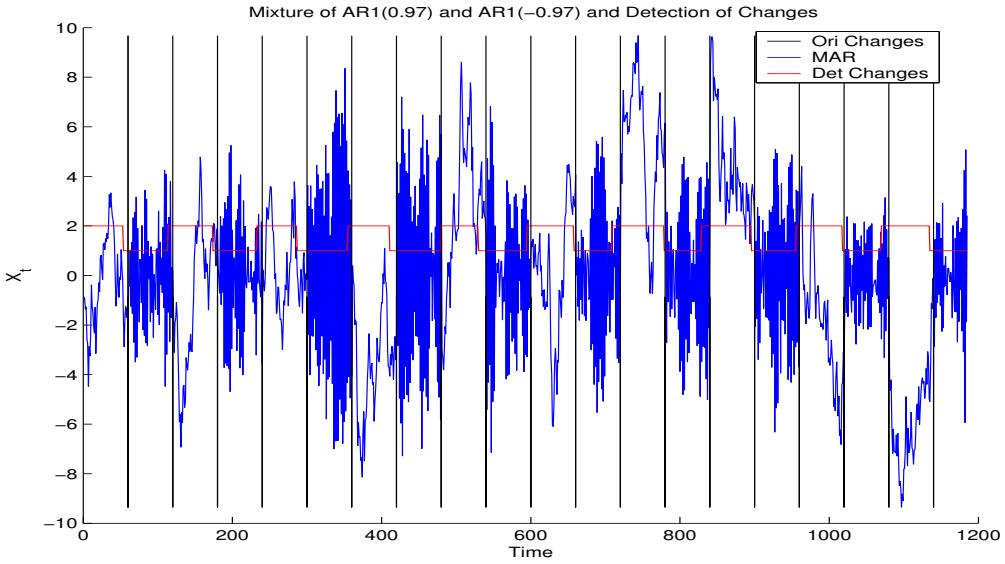

Figure 9.1: Detection of change: Mixture of AR(1)

The picture consists of a Mixture of AR processes(MAR) (the blue curve), the generated times of change that are represented by the vertical black lines and the estimated times of change represented by the jumps within the red curve.
The observed process was generated by a superposition of two AR(1) processes generated with autoregressive parameter $\alpha_1 = 0.97$ and $\alpha_2 = -0.97$. This superposition consists of taking the first 60 realizations of the mixture from the $AR1(.97)$ and the realizations 61 to 120 from the $AR1(-.97)$, in other words we consider $L = 60$ in the model defined in Equation 9.1-9.2. By doing so iteratively we obtain our desired process. For each underlying AR process the residuals were considered to be i.i.d. $\mathcal{N}(0,1)$.
The red curve estimates the presence in different states of the observed process and is also used to estimate the times of change. Indeed, the estimated times of change are represented by the different jumps within the red curve. This means that whenever a jump is observed, we estimate that we can move from one state to another one.

A close analysis of the above results conclude that up to some small perturbations, the times of change are nicely estimated on the training set and the validation set as well. Therefore, the weighted least squares techniques can be recommended for the segmentation of the type of mixture we presented above.

9.1.2 GMAR-ARCH(1) and Hidden Markov Techniques

In the previous section, we work with a regular and deterministic unknown time of change, an assumption that in many situations may be very far from approximating the reality. Therefore, in this section, we consider a GMAR-ARCH(1), i.e. a model of the form

$$X_t = \begin{cases} m_1(X_{t-1}) + \sigma_1(X_{t-1})\epsilon_t & \text{if} \quad S_t = 1 \\ m_2(X_{t-1}) + \sigma_2(X_{t-1})\zeta_t & \text{if} \quad S_t = 0, \end{cases} \tag{9.4}$$

where $S_t \in \{0, 1\}$ have to be considered as a first order Markov Chain with transition probability matrix A and the processes ϵ_t, ζ_t are independent sequences of i.i.d. $\mathcal{N}(0, 1)$ random variables.
Under our setting, we consider that the dynamic of the Markov Chain is driven by the transition probability matrix

$$A = \begin{pmatrix} 0.985 & 0.015 \\ 0.015 & 0.985 \end{pmatrix}$$

and we choose

$$m_1(x) = \alpha x + \beta e^{-\gamma(x-\mu)^2} \qquad m_2(x) = \frac{e^{\nu - x}}{1 + e^{\nu - x}}$$

and

$$\sigma_1(x) = \sqrt{\omega_1 + a_1 x^2} \qquad \sigma_1(x) = \sqrt{\omega_2 + a_2 x^2}.$$

Where $(\alpha, \beta, \gamma, \nu) \in \mathbb{R}^4, \omega_1 > 0, \omega_2 > 0$ and $a_1 \geq 0, a_2 \geq 0$.
Making use of the Markov structure, we generate 3000 realizations of the hidden process S_t which helps us to know which dynamics was used at each time instant for generating the observed process.
Unlike the theory presented in Chapter 4, for the generated data under the above model we assume the autoregressive and volatility functions can be suitably estimated by single layer feedforward networks. In this problem we are interested in the estimation of the autoregressive functions (at each time instant) and the times of change as well (the hidden process).
For this purpose, the usual technique for Neural Networks is taken into consideration, i.e. we split our data set into two subsets, namely the training set and the validation set. The first 1100 observations are used as training set and the remainder are used as validation set. The results of the simulation are presented in the following picture.

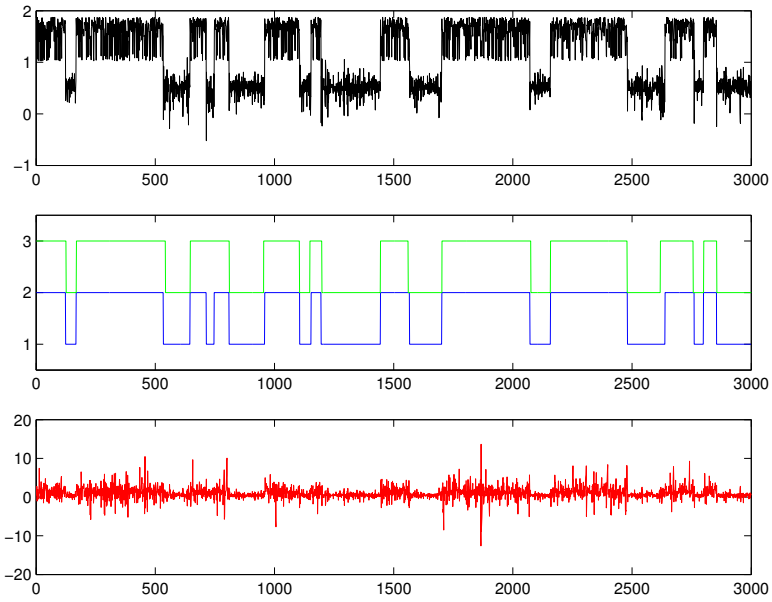

Figure 9.2: Estimations under Computer Generated GMAR-ARCH(1) model

The picture above is composed of three subplots. The first subplot that contains the black curve represents the estimated autoregressive part of the generated process at each time instant. The second subplot represents the generated hidden process and its estimates. In fact the green curve in this subplot represents $S_t + 2$, that is a shifted S_t generated according to the model in equation 9.4 and the blue represents $\hat{S}_t + 1$, where $\hat{S}_t$ is an estimate of S_t. The third subplot, i.e. the red curve represents the computer generated first order GMAR-ARCH data which are obtained via the generated hidden process.

For this estimation we consider exactly two states for the hidden Markov process as generated and use feedforward networks with 3 hidden layers for the autoregressive and volatility functions as well. Additionally, the $tanh$-functions were chosen as activation functions for the estimation of the autoregressive functions and the $logistic$ functions were considered for the estimation of the volatility functions.

As one can observe, on the training set, the hidden process is well estimated up to the perturbation that arises in a short period somewhere within the time span between 500 and 700 where the phase is not properly estimated. Hence, in general the statistical test for validating the changes will be of great help. Although there is no suitable test at the moment for validating the changes, the estimated hidden process fits well the generated hidden Markov process on the validation set.

9.2 Forecast of Daily Stock Values and Market Strategy

This section applies the Hidden Markov technique to the real-life data. The forecasting of the daily values of the DAX and those of BASF (one of its main components) is investigated. For this purpose, we first present the general situation and later on apply the model on each case with its particularities.

9.2.1 Model for Daily Stock Values

For the data we consider the daily stock values Y_t and transform them via a shifted logarithm, i.e. we define

$$X_t = \log Y_t - \log Y_1$$

for which we assume the following model

$$X_t = \sum_{k=1}^{3} S_{t,k}(m_k(\mathbb{X}_{t-1}) + \sigma_k(\mathbb{X}_{t-1})\varepsilon_t) \text{with } S_{t,k} = \begin{cases} 1 & \text{for } k = Q_t \\ 0 & \text{otherwise} \end{cases} \tag{9.5}$$

where $\{Q_t\}$ is assumed to be a stationary Markov Chain with values in $\{1, 2, 3\}$ and $\mathbb{X}_{t-1} = (X_{t-1}, \cdots, X_{t-p})$.
As previously, we assume that each of the autoregressive and volatility functions can be nicely approximated by a suitable feedforward network, that for sake of simplicity we choose to have the same amount of hidden neurons. Additionally, we choose the order of the autoregressive without any particular assumption. Under this consideration we are interested in the estimation of the autoregressive functions and the estimation of the hidden process that we assume to be a M.C. as announced previously. The estimate of the autoregressive function that we denote $\hat{f} : \mathbb{R}^p \longrightarrow \mathbb{R}$ is used to build up a Market Strategy that we will define after an application of the model in equation 9.5 to the forecast of transformed daily values of DAX and BASF respectively. In general the $tanh$ functions (they also return negative values) will account as activation functions for the estimation of the autoregressive function and the logistic functions (they only return positive values) are used as activation functions for the estimation of the volatility functions.

Forecast of Transformed Daily DAX Values

The data used in this section are downloaded from the Internet (http://www.markt-daten.de/daten/daten.htm), the original data represent the daily observed values of the DAX. We transformed the data according to the procedure defined earlier, that is considering X_t as defined in equation 9.5. We set $p = 3$ and make use of 5 hidden neurons for each autoregressive and volatility function. Last but not least, we recall that the hidden process is supposed to take values in 3 different states. Considering the first 1600 as training set we achieve the following results on the training set.

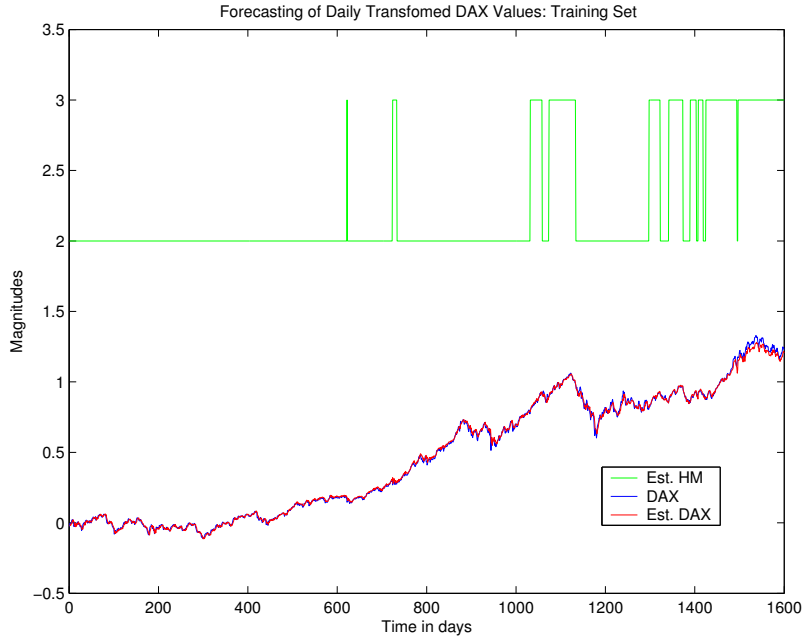

Figure 9.3: Forecast of Transformed Daily DAX Values: Training Set

And the following illustrates the related residuals.

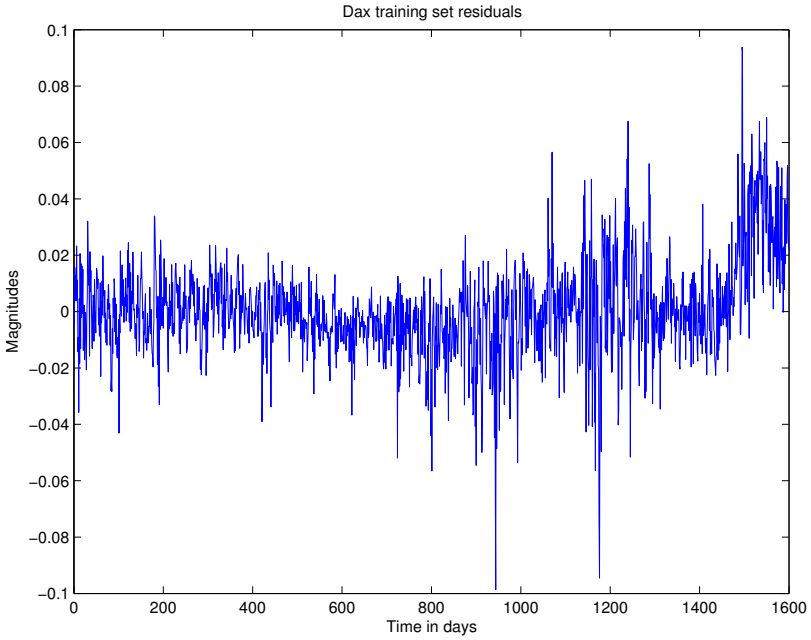

Figure 9.4: DAX residuals on the training set

In Figure 9.3, the green curve represents the estimated Hidden Markov process, the blue curve the first 1600 observation of the transformed DAX values from the 19^{th} January 1984 to the 4^{th} October 2004 and the red curve represents the estimated autoregressive functions from the model in equation 9.5.
From a first impression one has the feeling that the estimation procedure is in a certain sense not stable enough, but a deeper analysis of the Figure 9.3, e.g. zooming in the area of supposed instability delivers a completely different analysis. In fact, by zooming in on this area of apparent instability we observe that these perturbations account in general for changes over a short period, but in general not less than 10 days. Moreover, these short changes correspond to very volatile periods on the financial market. These short periods of change that help us to be closer to the reality of financial market may help us to avoid big losses over theses periods or even better may help us to achieve some gains in investment as it will be confirmed by the market strategy that we will present in the next section.
However, we have to remark that by assumption we have considered three states for the hidden process, but now the results exhibit only two of them. This may look surprising but the impression we have is that the procedure may be able to discard irrelevant dynamics.
We can observe that the estimated autoregressive functions approximated the transformed daily data well. However, we need to point out the presence of a small trend at the end of the training set on the residual plot. All these observations are confirmed by the validation set as the coming pictures illustrate.

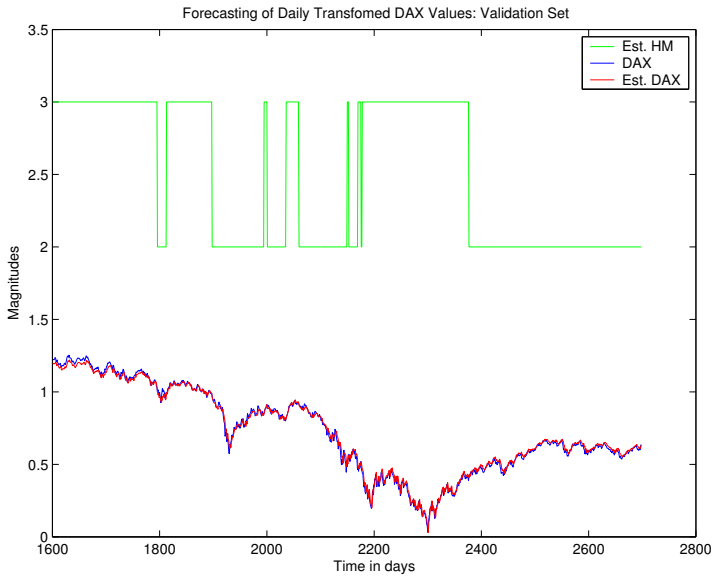

Figure 9.5: Forecast of Transformed Daily DAX Values: Validation Set

The related residuals plot is the following.

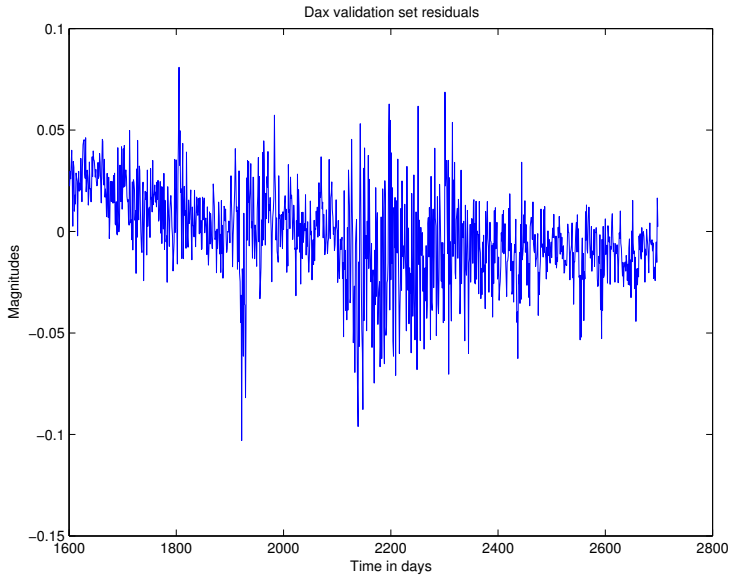

Figure 9.6: DAX residuals on the validation set

Unlike for the training set, in Figure 9.5 the green curve represents the estimated Hidden Markov process, the blue curve the remainder observations of the transformed DAX values from the 19^{th} January 1984 to the 4^{th} October 2004 and the red curve represents the estimated autoregressive functions from the model in equation 9.5 that we use to forecast the transformed daily DAX value, and in turn the daily DAX values

9.2.2 Forecast of Transformed Daily Values of a DAX Component: BASF

The data used in this section, like those of the DAX observations, were downloaded from Internet (http://www.corporate.basf.com/en/investor/aktie/kurs.htm) and they represent the transformed daily values of the BASF stock for the period from the 1^{st} July 1996 to the 12^{th} October 2004.
Once more we consider the transformed daily observations, in this case those of daily observed BASF values Y_t that we define as it follows

$$X_t = \log(Y_t) - \log(Y_1)$$

and assume a model as defined in equation 9.5 for X_t. For this case we take $p = 1$ and as previously, we assume that we can nicely approximate the autoregressive and volatility functions by suitable Feedfoward Networks, networks that we now assume to have 5 hidden neurons for each of the autoregressive and volatility functions. The

results based on the training set (the first 1000 observations) are illustrated by the following pictures.

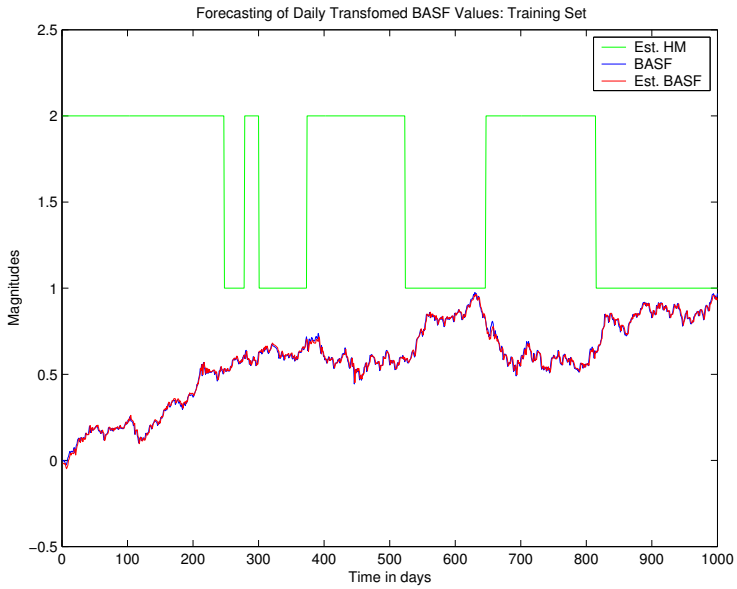

Figure 9.7: Forecast of Daily BASF Values: training Set

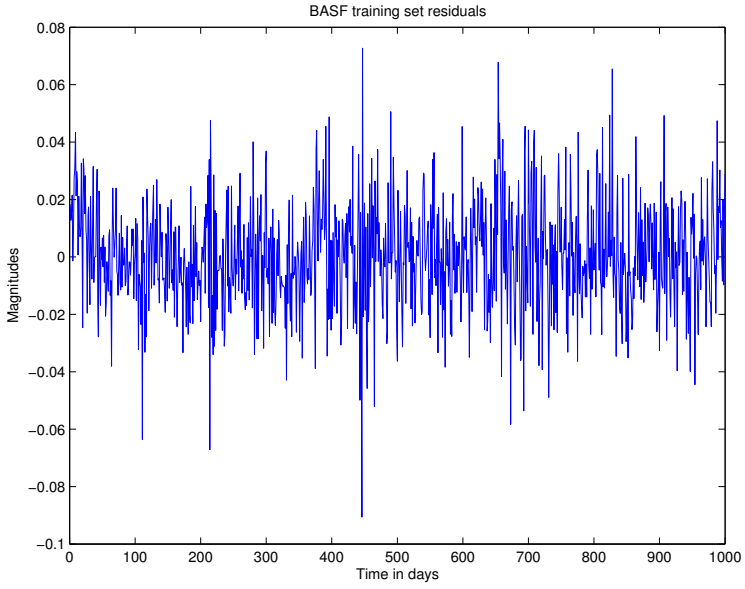

Figure 9.8: BASF residuals on the training set

The previous figure represents the residual plot and in Figure 9.7 the green curve

represents the estimated Hidden Markov driven process, the blue curve the observations of the transformed BASF values and the red curve represents the estimated autoregressive functions (at each time instant) from the model in equation 9.5.
Again, we observe on the training set that only two networks seem to be useful for our purpose. This observation is not completely verified on the validation set since the third network occurs on the validation set. This occurrence of the third network can be analyzed in many different ways, but what seems to be plausible here is a special role devoted to that network. Indeed, the point of occurrence of this network is located around the 11^{th} of September 2001. Therefore, one can say that although the third network is not present on the training it may be very useful for capturing extreme events.

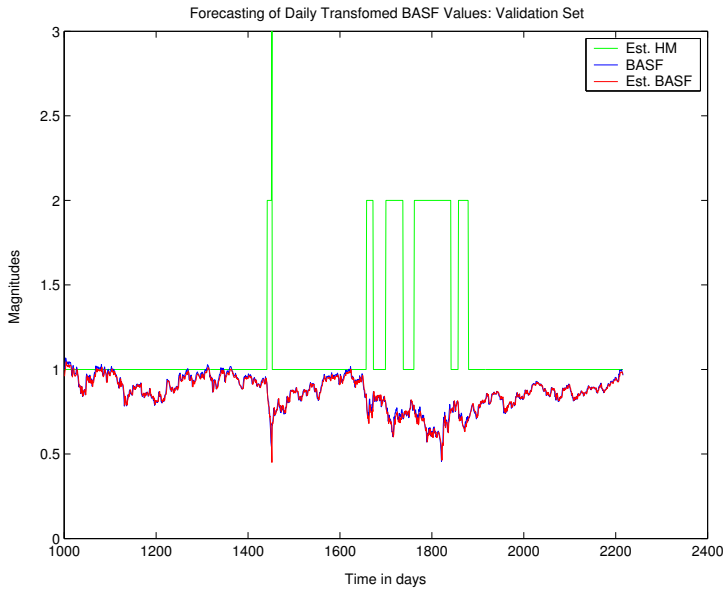

Figure 9.9: Forecast of Daily BASF Values: Validation Set

This extreme situation around the 11^{th} September is also depicted on the residuals plot where a big and unusual variation is observed as Figure 9.10 illustrates. However, there is no trend on the residuals plot compare to the DAX residuals plot. A question arises at this point, why does the third network act in the case of BASF and not in that of the DAX. An explanation may be the fact that the DAX cap is less homogenous than BASF. Indeed, there are some variations in the composition of the DAX cap over time, and therefore the difficulty in this case to predict extreme events.

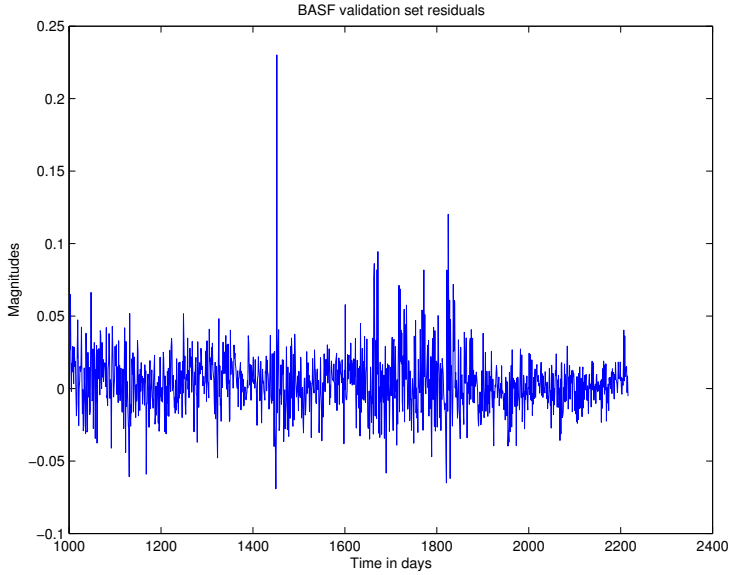

Figure 9.10: BASF residuals on the validation set

Although markets strategies are not central for the current project, in the next section, we present a basic market strategy build on the GMAR-ARCH model that we compare to the well-known Buy and Hold Strategy. In other words, we compare the investment results delivered by the signal getting from GMAR-ARCH strategy to that of a random walk based strategy.

9.2.3 Market Strategy

In this section, based on the forecast presented in the previous section we build up a market strategy that we compare to a strategy based on random walk model, i.e a buy and hold strategy.
Let us assume that we could have chosen a random walk model for the transformed data, i.e.

$$X_t = X_{t-1} + \varepsilon_t$$

with ε_t i.i.d. $\mathcal{N}(0, 1)$.
Then the value of today will have been considered as the best predictor of that of tomorrow. This just means that under a random walk model one can buy and hold, since the expected loss between today and tomorrow is equal to simply zero. Nevertheless, as one can observe using a buy and hold strategy is not the best one can do (in particular when there is a downward trend on the market). However, the trends are quite difficult to predict since the financial market act under various stochastic influences. Therefore, we need to consider alternative strategies that take into account the stochastic nature of the variations in the financial market.
As an alternative to the buy and hold strategy built on a random walk hypothesis,

based on the GMAR-ARCH model we proposed a simple strategy that is based on the following simple idea:
If there is a positive signal (what we will define next), one should buy the stock in consideration with all the money available at that time in the saving account or hold the stock one already has in possession at that time. Conversely, if there is a negative signal one should not buy any stock, moreover one should sell all the stocks in possession at that time or keep the money in the saving account if one does not possess any stock at that time.
For this strategy we make the following assumption

A. 9.1 *(Main Assumption)*
There is no transaction cost and there is no interest rate on saving accounts.

Under the above hypothesis, let us now consider Y_t to be the observed process, e.g. the daily observed DAX values and X_t the transformed DAX values. Under our considerations,

$$X_t = \log(Y_t) - \log(Y_1)$$

and X_t follows a GMAR-ARCH model for which we assume that the autoregressive functions can nicely be estimated at each time point by suitable neural networks that we denote $\hat{f}(X_{t-1}, \cdots, X_{t-p})$. The announced trading strategy is then built as follows.
Given an initial value of the observed series Y_0 and an initial wealth G_0 that we are to invest in a given stock we need to define the wealth process G_t, that is the amount of money related to the given stock given the initial wealth. The wealth process is built on the auxiliary process Am_t which is the corresponding amount of stock given the signal.
At each time instant t we compute $\hat{f}(X_t, \cdots, X_{t+1-p}) - X_t$ (the announced signal) and define

$$Am_t = \begin{cases} \frac{G_t}{Y_t} & \text{if} \quad \hat{f}(X_t, \cdots, X_{t+1-p}) - X_t > 0 \\ 0 & \text{otherwise} \end{cases}$$

and the decision for the next time is made as it follows

$$G_{t+1} = \begin{cases} Am_t Y_{t+1} & \text{if} \quad Am_t > 0 \\ G_t & \text{if} \quad Am_t = 0. \end{cases}$$

This trading strategy was applied on the validation set for the DAX and BASF values with initial wealth of 1000 units. Moreover the wealth processes generated in both cases are compared to those generated by the Buy and Hold strategy. The results are presented in the coming pictures where in both cases the blue curve represents the wealth process generated by the Hidden Markov based strategy and the red curve is the wealth process as delivered by the Buy and Hold strategy. In both situations, the hidden Markov based strategy performs better than the Buy and Hold strategy.

Additionally, given the initial wealth we never make any loss with the first strategy. Moreover, the first strategy performs better for the BASF values compare to the DAX values.

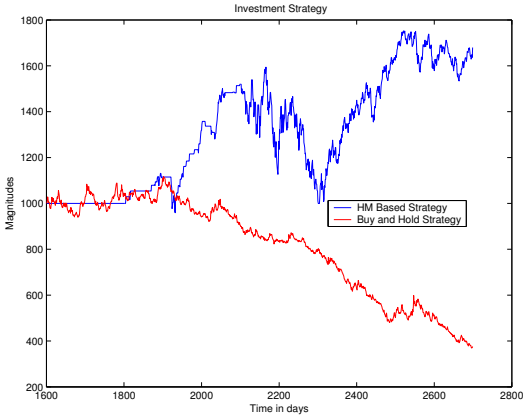

Figure 9.11: Strategies for DAX Values

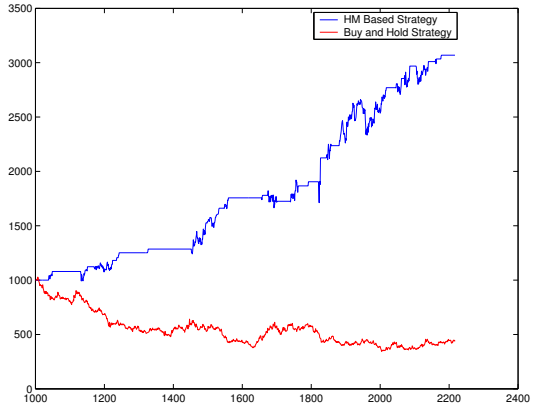

Figure 9.12: Strategies for BASF Values

9.3 GMAR-ARCH as Model for DAX Return

In this section we consider a model of the type defined in equation 9.5 for the log-return of the DAX, i.e.

$$X_t = \log\left(\frac{Y_t}{Y_{t-1}}\right),$$

where as previously the X_t are daily observed DAX values.
Under this setting, we make use of 2500 observations that account for the period from November 1994 to October 2004. We take the autoregressive order $p = 3$ like for the forecast of the transformed daily DAX values. We also make the usual assumption on the estimation of the autoregressive and volatility functions by suitable single layer networks that we consider here to have each 6 hidden neurons. We use the first 1400 as training set and the remainder as validation set. Under this setting we are interested by the estimation of the autoregressive part of the model at each time instant and the related hidden observation.
Unlike for the daily DAX values, we have considered three states for the hidden process, but the difference here is due to the fact that all these states are duplicated in training and validation set as well. Moreover the interpretation of the different phases seems to be obvious (we will detail next) by a close observation of the plot of return within the period under consideration. Before we make any further comment let us first present the numerical results. The following picture consists of three curves, the red curve represents the log-returns, the black their autoregressive estimates and the blue curve is the estimated hidden process.

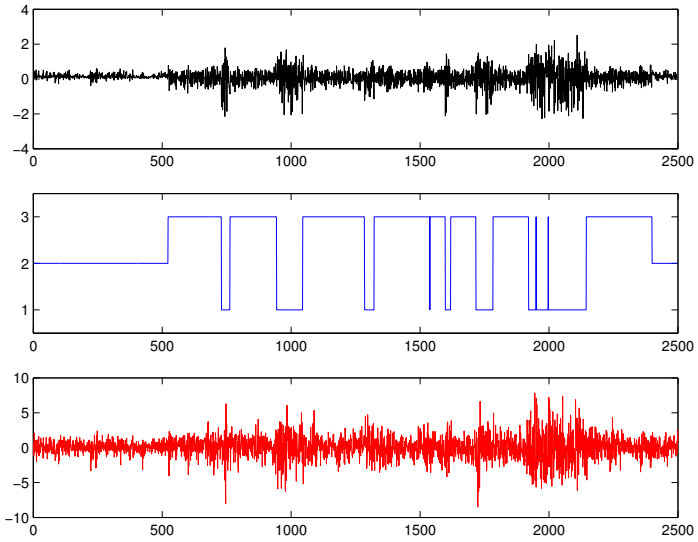

Figure 9.13: Return Estimation under GMAR-ARCH Model

The second network appears on the training set essentially at the beginning and on the validation right at the end, these two periods correspond to relative quietness on the financial market. The first network seems to model the highly volatile phases of the market as one can observe on the training and validation sets. Last but not least, one can say that the third network models the phases intermediate phases between the previous two phases.
As one can observe, we can use the model to have a nice segmentation of the market phases. However, we need to point out the problem we have with the estimation of the autoregressive part, in some numerical results, they were phases where the autoregressive part was either always positive or always negative. Hence, the difficulty to use the autoregressive part in those situations.

10 Conclusion and Outlook

In this chapter we first summarize what has been done in this thesis, exhibit some fields of further investigation.

10.1 Conclusion

In current work, many problems of interest are addressed. We can mention the extension of Nonlinear AR-ARCH models to the Generalize Mixture of AR-ARCH as defined in Chapter 2. For this model we provide some sufficient conditions to ensure the asymptotic stability of the system, these conditions are fulfilled under various regularity assumptions.
Given a special consideration of the autoregressive functions in this model, we establish the universal approximation (consistency of function estimates) property of these functions by some parametric classes of functions and particularly some classes of single layer Feedforward Networks. Further, considering both autoregressive and volatility functions, based on the hidden structure, we define a conditional likelihood from which we derive a pseudo conditional $\log$-likelihood under normality assumption of the residuals and suitable approximation of the autoregressive and volatility functions by single layer Feedforward Networks. For the pseudo conditional $\log$-likelihood we prove the consistency of the parameter estimates and design a version of the well-known EM algorithm for the numerical extraction of the model parameters.
Beside this Hidden Markov Driven models, we also consider a nonlinear weighted least squares approach for estimating the time of change and the autoregressive functions at each instant. Under analogous consideration as previously, i.e. we consider Feedfoward Networks as function estimates and prove the consistency and asymptotic normality of the parameter estimates. Unfortunately, here we consider a mixture of independent stationary processes.
Practically, one can say that the weighted least squares techniques and the Hidden Markov Driven approach perform quite well as some case studies in Chapter 9 illustrate.
Despite these successful applications of both approaches, it is important to mention the necessity of suitable statistical tests for validating the changes in the dynamics of the observed process. In this light, we consider a special case (compare Chapter 8) for which we build up a statistical test based on the moving sum technique introduced by Huskova [52].
However, it is also important to mention that many questions, which merit further considerations, arise along the way and we will recall some of the most important ones in the next section. Additionally, we suggest an alternative formulation of the type of problem solved in this thesis.

10.2 Outlook

Along the chapters of this thesis we have mentioned several unsolved problems of particular importance. Among these problems, we can recall that of the the choice of the order of the autoregressive and volatility functions. In other words we need to build some information criterion that may help for the order selection. The order selection is also closely related to the choice of the order of the Hidden Markov process and the number K of states of this hidden process. Additionally, considering the approximation of autoregressive and volatility by suitable neural networks, the problem of the choice of the number of hidden neurons can be pointed out.
Beside the problems related to information criterion, the problem of validating the changes is crucial. This problem is partially addressed in Chapter 8 where a mixture of nonlinear AR process with only one abrupt unknown change is considered. The challenge is to develop similar tests for Generalized Mixture of AR-ARCH models and to study the asymptotic property of the estimated time of change, which in the case mentioned in Chapter 8 can be considered as a byproduct of the developed test. An interesting issue is also to extend all the above considerations to a multivariate time series, in turn to a portfolio. Furthermore, one can focus on the hidden process for which one can consider a long memory process compare to a first order Markov Chain. Under such consideration one will first like to extend our theory to cover this type of situation or depending on the applications one may also solve the problem in terms of occupation time or number of jumps from one state to another one.

A An Introduction to Neural Networks

A.1 Preliminaries and Network Description

The term Neural Network has evolved along the time and nowadays represents a very large class of models and learning methods. Its comparison to the human brain makes it mysterious, however it is still too far from describing the reality of the human brain. As it will be made clear in this chapter, a Neural Network can simply be regarded as a parametric nonlinear statistical model, much like projection pursuit among others.
Typically a Neural Network has a graphical representation as in figureA.1 where the $(X_1, \cdots, X_p)$ represent the Input variables, the $(Y_1, \cdots, X_q)$ the Output variables and(in between) the $(Z_1, \cdots, Z_H)$ represents the Hidden layers. The latter are so called because there are not directly observable.

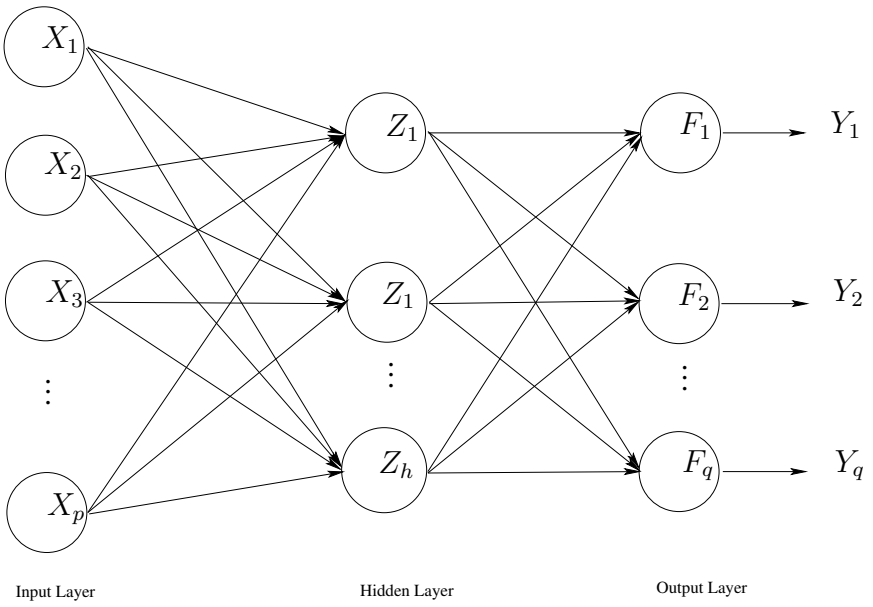

Figure A.1: Feedfoward Network

In this chapter we focus on the most widely-used class of Neural Network, namely on the class of single Layer Feedfoward Network. For sake of simplicity we consider $q = 1$, i.e., we consider that the output to represent a single variable. Nevertheless one can refer to the book by Haykin [46] for a more general introduction into this area or to the book by Anders [2] for an introduction and application to the Credit Scoring.
Under our setting we have to recall that (as usual), the hidden features Z_h are derived from linear combinations of the input variables and the target Y is defined as

a function of the linear combinations of the Z_h, that is

$$\begin{aligned} Z_h &= \varphi_h(\alpha_{0,h} + \alpha_{1,h}X_1 + \cdots + \alpha_{p,h}X_p) \\ &= \varphi_h(\alpha_{0,h} + \alpha_h' X) \quad \text{for } h = 1, \cdots, H \end{aligned} \tag{A.1}$$

and

$$Y = F_H(\nu_0 + \nu_1 Z_1 + \cdots + \nu_H Z_H) = \nu_0 + \sum_{h=1}^{H} \nu_h Z_h. \tag{A.2}$$

The functions φ_h, which may be different for different values of h are called activation functions and often taken to be the type of sigmoid with the most popular choice being the sigmoid function, i.e.

$$\varphi_h(x) = \frac{1}{1 + e^{-x}}$$

and F_H is a given function for which the identity function or the truncated identity function can always be considered to solve regression problems.

A.1.1 Some Examples of Activation Functions

In Neural Network literature one can find various type of activation functions, in this section we recall few of them. For more detail knowledge on this issue one can refer to the literature mentioned previously in this chapter.

1. Linear functions of the form $\varphi(x) = \alpha x$(for some $\alpha \geq 0$). Considering this activation function for every hidden unit, the entire system collapses to a linear model. Therefore, one can consider Neural Networks as nonlinear generalization of linear models, e.g., for regression problems.

2. Threshold function that is defined as

$$f(x) = \begin{cases} 1 & \text{if} \quad x > a \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

   where a is given bound. The next picture represents such function for $a = 1$.

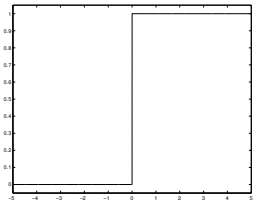

Figure A.2: Threshold Activation function

3. Piecewise linear functions that for some $\alpha > 0$ are defined as

$$f(x) = \begin{cases} 0 & \text{if} \quad x < -\alpha \\ \frac{1}{2}(\frac{x}{\alpha} + 1) & \text{if} \quad -\alpha \leq x \leq \alpha \\ 1 & \text{if} \quad \alpha < x \end{cases}$$

In the following picture we give a graphical representation of such function with $\alpha = 1$.

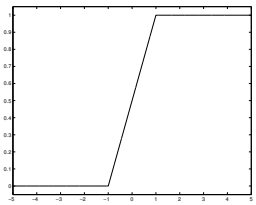

Figure A.3: Piecewise Linear Activation Function

4. Sigmoid Function. This represents the most common activation used in the construction of neural networks. It is defined as a strictly increasing function that exhibits a graceful balance between linear and nonlinear behavior. An example is the well-known logistic function

$$\varphi(x) = \frac{1}{1 + \exp(ax)}, \tag{A.4}$$

where a is the slope parameter of the sigmoid function. For the latter class of activation function, taking for example a big enough compare to 1 will account for a hard thresholding as one can observe from the following picture.

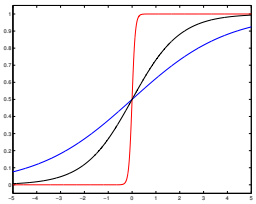

Figure A.4: Type of Sigmoid Activation Functions

In the above picture we have different representations of $\varphi(x) = \frac{1}{1+\exp(-ax)}$. For the black curve $a = 1$, for the red curve $a = 15$ and last but not the least we consider $a = 0.2$ for the blue curve.

It may also be necessary to allow the activation function to take on negative values. In such situations the use of hyperbolic tangent functions, defined by

$$\varphi(x) = \tanh(x) \tag{A.5}$$

may be highly recommended.

A.2 Neural Networks in Practice

A.2.1 Least Squares

The Neural Networks have unknown parameters (usually called weights) which we want to estimate, i.e. we have to find the values of the weights that make the model fit the data well. For this purpose we need to be given a training set

$$(X_t, Y_t), \quad t = 1, \cdots, n$$

with $X_t \in \mathbb{R}^p,\ Y_t \in \mathbb{R}^p$ and for the regression problem we can use the sum of square error

$$Q(\theta) = \sum_{t=1}^{n} (Y_t - f_H(X_t))^2 = \sum_{t=1}^{n} Q_t,$$

where

$$f_H(X_t) = \nu_0 + \sum_{h=1}^{H} \nu_h \varphi(\alpha_{0,h} + \alpha_h' X_t).$$

For sake of simplicity we consider the same activation for all hidden units; $\theta \in \mathbb{R}^{H(p+2)+1}$ consists of

$$\nu_0, \nu_h, h = 1, \cdots, H; \quad \alpha_{0,h}, \alpha_{i,h}, i = 1, \cdots, p,\ h = 1, \cdots, H.$$

A.2.2 Backpropagation

The standard approach used for the minimization of $Q(\theta)$ is a stochastic approximation that we call Backpropagation in our context. Backpropagation is so popular because it is easy to implement. However, it can suffer from several drawbacks that we will discuss later.
The algorithm can simply be defined as follows. Starting with a weight $\hat{\theta}_0$ one derives the weight in the next step by the recursion formula

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \gamma_n \nabla Q(\hat{\theta}_n),$$

where ∇Q is the vector of first partial derivatives of Q with respect to θ and γ_n is the learning that satisfies some technical conditions, namely

$$\gamma_n \longrightarrow 0 \text{ as } n \longrightarrow \infty, \sum_n \gamma_n = \infty, \text{ and } \sum_n \gamma_n^2 < \infty.$$

These conditions are satisfied for $\gamma_n = \frac{1}{n}$. Therefore, one can observe that Backpropagation is a form of the stochastic gradient algorithm due to Robbins and Monro(1951).

Drawback of Backpropagation and Alternative Solutions

The main drawback of backpropagation is the fact that if the dimension of the parameter is very high or the size of the training set is large enough the algorithm may become very slow and for this reason one would rather rely on second order techniques such as the Newton algorithm. But the latter algorithm is not the best one can use, in the sense that in this case one will need an explicit computation of the matrix of second derivatives, which can be very large in this case . To overcome this problem one can use a conjugated gradient algorithm or a variable metric method, both approaches avoid explicit computation of the Hessian while providing faster convergence.

A.3 Some Technical Remarks

A.3.1 Input

It is recommended to scale all the input around their sample mean and variance. This will ensure that all input are treated almost equally and may improve the quality of the results. In this situation we can easily take our starting value as uniformly distributed on the interval $[-a, a]$. However, if we choose a to be too close to zero, this just means that we allow our model to start into a linear model because in this case the neural network collapses into a nearly linear model. As the weights increase, we go back to a nonlinear estimation problem.

A.3.2 Local minima

The objective functions are by nature non convex, therefore possessing many local minima. As a consequence, the final result obtained is always dependent on the choice of starting value. Therefore we will at least try a number of random starting points and choose the solution given the best solution to our problem. Since we are in a nonlinear setting this approach is better than averaging over all the final solutions.

A.3.3 Number of Hidden Neurons

Roughly speaking it is better to have too many neurons than too few. With too few hidden neurons, the model might not provide enough flexibility to capture all nonlinear aspects in the data. With too many neurons, extra weights can be set to zero by an appropriate regularization.

References

[1] Takeshi Amemiya. *Advanced Econometrics* . Havard University Press, USA, 1985.

[2] Ulrich Anders. *Statistische Neuronale Netze*. Verlag Franz Vahlen GmbH, München, 1997.

[3] Leonard E. Baum, Ted Petrie, Georges Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1), 1970.

[4] Herman J. Bierens. *Robust methods and asymptotic theory*. Lecture notes in economics and mathematical sytem 192, Springer Verlag, Berlin, 1981.

[5] Herman J. Bierens. Sample moments integrating normal kernel estimators of density and regression functions. *Sankhya*, 45, 1983.

[6] Patrick Billingsley. *Statistical Inference for Markov Processes* . The university of Chicago Press, USA, 1961.

[7] Patrick Billingsley. *Ergodic Theory and Information* . John Wiley & Sons, New York, USA, 1965.

[8] Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, USA, 1968.

[9] Patrick Billingsley. *Probability and Measure* . John Wiley & Sons, New York, Usa, 1986.

[10] Denis Bosq. *Nonparametric Statistics for Stochastic Processes*. Lecture Notes in Statistics 110. Springer-Verlag, Heidelberg, 1998.

[11] Leo Breiman. *Probability* . Addison-Wesley, USA, 1968.

[12] Peter J. Brocwell and R A. Davis. *Time Series: Theory and Methods* . Springer-Verlag, Germany, 1991.

[13] Myklos Csörgo and L Horvath. *Limit Theorems in Change-point Analysis* . John Willey & Sons, USA, 1997.

[14] Y. Davydov. Mixing conditions for markov chain. *Theory of probability and its applications*, 18(2), 1973.

[15] Georgi Dimitroff. *Neural Network-based estimates for volatility*. Diplom Thesis, University of Kaiserslautern, 2001.

[16] Jürgen Dippon. Asymptotic analysis of a learning method in neural network. *Tagungsbericht Neuronale Netze in Ingenieuranwendungen (Hg. B. Krplin). ISD Univ. Stuttgart*, 1996.

[17] Vincent Dortet-Bernadet. Choix de models pour chaîne de markov cachées. *Académie des sciences, Editions scientifiques et médicales*, 332, 1989.

[18] Paul Doukhan. *Mixing - Properties and Examples*. Lecture Notes in Statistics 85. Springer-Verlag, Heidelberg, 1994.

[19] Paul Doukhan and M. Ghindes. Etude du processus $X_{n+1} = f(X_n) + e_n$. *C.R.A.S. Série A*, 290, 1980.

[20] Marie Duflo. *Random Iterative Models*. Springer-Verlag, Heidelberg, 1997.

[21] Ernst Eberlein and Murad S. Taqqu. *Dependence in Probability and Statistics: A Survey of Recent Results* . Birkhaeser, Germany, 1986.

[22] Robert J. Elliot, Lakhdar Aggoun, and John B. Moore. *Hidden Markov Models: Estimation and Control* . Springer, USA, 1995.

[23] Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50, 1982.

[24] Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4), 1968.

[25] Jürgen Franke. Nonlinear and nonparametric methods for analyzing financial time series. *In P. Kall and H.-J. Luethi, eds., Operation Research Proceedings 98. Springer-Verlag, Heidelberg*, 1998.

[26] Jürgen Franke. Portfolio management and market risk quantification using neural networks. *In W.S. Chan, W.K. Li and H. Tong, eds., Statistics and Finance: An Interface. Imperial College Press, London* , 2000.

[27] Jürgen Franke and Mabouba Diagne. Estimating market risk with neural networks. *working paper*, 2001.

[28] Jürgen Franke, Wolfgang Härdle, and Christian Hafner. *Statistik der Finanzmärkte* . Springer-Verlag, Heidelberg, 2001.

[29] Jürgen Franke and Gerald Kroisandt. Nonparametric changepoint detection for time series. *To be published in Statistics and Decision*, 2004.

[30] Jürgen Franke, Michael H. Neumann, and Jean P. Stockis. Bootstrapping nonparametric estimates of the volatility function. *Report in Wirtschaftsmathematik , University of Kaiserslautern. To appear in Journal of Econometrics* , 77, 2001.

[31] Andrew M. Fraser and Alexis Dimitriadis. Forecasting probability densities by using hidden markov models with mixed states,. *in A. Weigend and N. Gershenfeld(eds), Time Series Prediction: Forecasting the Future and Understanding the Past, Addison Wesley*, 1993.

[32] A. Ronald Gallant. *Nonlinear Statistical Models*. John Wiley & Sons, New York, 1987.

[33] Valentine Genon-Catalot, T Jeantheau, and C Laredo. Conditional likelihood estimators for hidden markov models and stochastic volatility models. *Scandinavian journal of statistics*, 30, 2003.

[34] Walter R. Gilks, Sylvia Richardson, and David J. Spielgelhalter. *Markov Chain Monte Carlo in Practice* . Chapman and Hall, England, 1996.

[35] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *Der Latex-Begleiter*. Addison Wesley, München, Germany, 2000.

[36] Christian Gourieroux. *ARCH Model and Financial Application*. Springer-Verlag, New York, 1997.

[37] Ulf Grenander. *Abstract inference*. Wiley, New York, 1981.

[38] Dominique Guégan and Jean Diebolt. Probabilistic properties of the β-arch model. *Statistica Sinica*, 4, 1994.

[39] Lszl Györfy, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression* . Springer-Verlag, Heidelberg, 2002.

[40] Christian Hafner. *Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility*. Physica-Verlag, Heidelberg, 1998.

[41] James D. Hamilton. Rational expectations econometric analysis of changes in regimes: An investigation of the term structure of inetrest rate. *Journal of Economic Dynamics and Control*, 12, 1988.

[42] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 1989.

[43] James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, 1994.

[44] Edward J. Hannan. *Multiple time series*. Wiley, New York, 1970.

[45] Trevor Hastie, Robert Tibshirani, and Jerome Friedmaan. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, New York, Berlin, Heidelberg, 2001.

[46] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, 1999.

[47] Ulla Holst, Georg Lindgren, Jan Holst, and Mikael Thuvesholmen. Recursive estimation in switching autoregressive with a markov regime. *Journal of time series analysis*, 15(5), 1994.

[48] Kurt Hornik. Approximation capabilities of multilayer feedfoward networks. *Neural Networks*, 4, 1991.

[49] Kurt Hornik, M Stinchcombe, and Halbert White. Multilayer feedfoward networks are universal approximators. *Neural Networks*, 2, 1989.

[50] Lajos Horváth, Piotr Kokoszka, and Gilles Teyssière. Empirical process of the square residual of an arch sequence. *The Annals of Statistics*, 29(2), 2001.

[51] Peter J. Huber. *Robust Statistics*. John Wiley and Sons, New York, Berlin, Heidelberg, 1981.

[52] M. Huskova. Recursive m-test for detection of changes. *Sequetial analysis*, 7, 1988.

[53] Marie Huskova and Jaromir Antoch. Detection of structural changes in regression. *Tatra Mountains: Mathematical Publications*, 26, 2003.

[54] Sylvia Kaufmann and Sylvia Frühwirth-Schnatter. Bayesian analysis of switching arch models. *World Conference Econometric Society, Seattle*, 2000.

[55] Lawrence A. Klimko and Paul I. Nelson. On conditional least square estimation for stochastic processes. *The Annals of statistics*, 6(3), 1978.

[56] Ulrich Krengel. *Ergodic Theorems* . Walter de Gruyter, Berlin, New York, 1985.

[57] Gerald Kroisandt. *Change-point Analysis with Wavelets for Time Series with Structural Jumps* . Phd Thesis Kaiserslautern University, Germany, 1998.

[58] Harold J. Kushner and Dean Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York, Heidelberg, Berlin, 1978.

[59] Lehnart Ljung and T Söderström. *Theory and practice of Recursive Identification*. The MIT Press, England, 1983.

[60] Lennart Ljung. Analysis of recursive stochastic algorithm. *IEEE Trans. Automatic Control*, AC-22, 1977.

[61] Lennart Ljung. Strong convergence of stochastic approximation algorithm. *Annals of Statistics*, 6(3), 1978.

[62] Lennart Ljung, Georg Pflug, and Harro Walk. *Stochastic Approximation and Optimization of Random Systems*. Birkhäser Verlag, Basel, Boston Berlin, 1992.

[63] Zudi Lu. On the geometric ergodicity of a non linear autoregressive model with an autoregressive conditional heteroscedastic term. *Statistica Sinica*, 8, 1998.

[64] Iain L. MacDonald and Walter Zucchini. *Hidden Markov and Other Models for Discrete-valued Time Series* . Chapman & Hall, England, 1997.

[65] Rachel J. MacKAY. Estimating the order of a hidden markov model. *The Canadian journal of Statistics*, 30(4), 2002.

[66] Gisela Maercker. Efficient estimation in ar models with arch errors. *www.mathematik.tu-bs.de/preprints*, 1996.

[67] Elias Masry and Dag Tjostheim. Nonparametric estimation and identification of nonlinear arch time series: Strong convergence and asymptotic normality. *Econometric theory*, 11, 1995.

[68] R. Meir. Nonparametric time series prediction through adaptive model selection. *Machine Learning*, 39, 2000.

[69] M. Métivier and P. Priouret. Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant. *Probability Theory and related fields*, 74, 1987.

[70] Sean P. Meyn and Richard L. Tweedie. *Markov Chain and Stochastic Stability*. Springer-Verlag, London, 1993.

[71] Dharmendra S. Modha and Elias Masry. Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inf. Theory*, 42, 1996.

[72] Dharmendra S. Modha and Elias Masry. Memory-universal prediction of stationary random processes. *IEEE Trans. Inf. Theory*, 44, 1998.

[73] Klaus R. Müller, Jens Kohlmorgen, Klaus Pawelzik, and Non-members. Analysis of switching dynamics with competing neural networks. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E78-A*, 10, 1995.

[74] M B. Nevelson and Hasminskii. *Translation of Mathematical Monographs volume 47: Stochastic Approximation and Recursive Estimation* . American Mathematical Society Providence, USA, 1973.

[75] A B. Poritz. Hidden markov models: A guide tour. *In International Conference on Acoustic, Speech and Signal Processing*, 7, 1988.

[76] Benedikt M. Pötscher and Ingmar R. Prucha. *Dynamic Nonlinear Econometric Models: Asymptotic Theory* . Springer, Germany, 1997.

[77] M B. Priestley. *Non-linear and Non-Stationary Time Series Analysis* . Academic Press, England, 1991.

[78] Richard E. Quandt. A new approach of estimating switching regressions. *Journal of the American Statistical Association*, 67, 1972.

[79] Lawrence R. Rabiner. A tutorial on hidden markov models and selected application in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.

[80] D. Revuz. *Markov Chain*. North-Holland International Library, Holland, 1984.

[81] Jean Schmets. *Theorie de La Mesure: Note du Cours d'analyse superieure de la premiere licence en science mathematiques*. Université de Liege, Faculté des Sciences, 1989-1990.

[82] Laurent Schwartz. *Theorie des Ensembles et Topologie* . Hermann, France, 1991.

[83] Jean P. Stockis and Jürgen Franke. Mixture of kernel estimates. *Working paper, University of Kaiserslautern*, 2003.

[84] Howell Tong. *Nonlinear Time Series: A Dynamical System Approach* . Oxford University Press, Oxford, 1990.

[85] H. Walk. An invariance principle for the robbins-monro process in a hilbert space. *Wahrscheinlichkeitstheorie und verwandte Gebiete*, 39, 1977.

[86] Andreas S. Weigend and Shanming Shi. Predicting daily probability distribution of S&P500 returns. *Journal of Forecasting*, 19, 2000.

[87] Halbert White. *Asymptotic Theory for Econometricians*. Academic Press, New York, 1984.

[88] Halbert White. Some asymptotic results for learning in single hidden-layer feedfoward networks models. *Journal of the American Statistical Association*, 84, 1989.

[89] Halbert White. Connectionist nonparametric regression: Multilayer feed-foward networks can learn arbitrary mappings. *Neural Networks*, 3, 1990.

[90] Halbert White and Iain Domowitz. Nonlinear regression with dependent observations. *Econometrica*, 52(1), 1984.

[91] Halbert White and et al. *Artificial Neural Networks: Approximation and Learning Theory*. Blackwell, Oxford, Cambridge, 1992.

[92] Halbert White and Jeffrey M. Wooldridge. Some results for sieve estimation with dependent observations. *In W. Barnett, J. Powell and G. Tauchen, eds., Nonparametric and Semi-Parametric Methods in Econometrics and Statistics. Cambridge University Press* , 1990.

[93] David Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991.

[94] Chun S. Wong and Wai K. Li. On mixture of autoregressive model. *Journal of royal statistical society*, 62(1), 2000.

[95] Chun S. Wong and Wai K. Li. On a logistic mixture of autoregressive model. *Biometrika*, 88(3), 2001.

[96] Chun S. Wong and Wai K. Li. On mixture of autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association*, 96(455), 2001.

[97] Jian F. Yao. On least squares estimation for stable nonlinear ar processes. *Ann. Inst. Stat. Math.*, 52(2), 2000.

[98] Jian F. Yao and J.-G. Attali. On stability of nonlinear ar processes with markov switching. *Apply Probability Trust*, 32, 2000.

Some Biographic Notes

Name, Date and Place of Birth

Tadjuidje Kamgaing, Joseph
Date of Birth 13.05.1973
Place of Birth Mbo, Cameroon

Education

- **Academics**
  - **Since November 2001,**
    PhD candidate at the department of mathematics of the university of Kaiserslautern in collaboration with Fraunhofer Institut für Techno und Wirtschaftsmathematik, Germany
  - **October 1999 to October 2001**
    Master program in Financial Mathematics at the university of Kaiserslautern, Germany.
  - **January to July 1999**
    German courses in Frankfurt and Kaiserslautern, Germany
  - **December 1998**
    Arrival in Germany
  - **1993–1998**
    Study of mathematics at the university of Yaoundé 1, Cameroon
- **School**
  - **1985–1993**
    Secondary school in Bafoussam (Government Bilingual Secondary School and Lycée Classique), Cameroon
  - **1979–1985**
    Primary school in Bafoussam (Ecole Annexe 1 B), Cameroon